THE ANNALS
of
MATHEMATICAL STATISTICS

The Annals of Mathematical Statistics Is Affiliated with the American Statistical Association and Is Devoted to the Theory and Application of Mathematical Statistics

Volume VII, Number 4
DECEMBER, 1936

PUBLISHED QUARTERLY
ANN ARBOR, MICHIGAN

Four Dollars per annum

Made in United States of America

Address: ANNALS OF MATHEMATICAL STATISTICS
Post Office Box 171, Ann Arbor, Michigan

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD.

ON A GENERAL SOLUTION FOR THE PARAMETERS OF ANY FUNCTION WITH APPLICATION TO THE THEORY OF ORGANIC GROWTH

By Harry Sylvester Will

Part I

I. **The Problem Stated.** A type of problem which continually arises in the ordinary course of statistical analysis is that of determining the numerical values of the parameters of a function used to represent a series of observational data. In mathematical terminology, the problem may be stated as follows:

Given, the observational series $Y_0, Y_1, \cdots Y_{n-1}$.

Assumed, the function $y = f(x, a, b, c, \cdots)$.

To find, the numerical values of the parameters $a, b, c, \cdots$.

If the function $f(x, a, b, c, \cdots)$ is linear in the parameters, the desired solution is easily obtained by familiar methods. In cases where the function is not linear, the standard procedure is to reduce it to the linear form by expansion into Taylor's series, thus:

$$f(x, a, b, c) = f(x, a_0b_0c_0) + f_a(x, a_0b_0c_0)\cdot\Delta a + f_b(x, a_0b_0c_0)\cdot\Delta b + f_c(x, a_0b_0c_0)\cdot\Delta c, \tag{1}$$

where $a = a_0 + \Delta a$, $b = b_0 + \Delta b$, $c = c_0 + \Delta c$.

The use of this method suffers from the excessive labor involved as the number of parameters to be determined increases. In cases where satisfactory values of the first approximations $a_0b_0c_0$ are not obtainable, the solution becomes impossible. The basic difficulty arises from the consideration that the Taylor theorem requires that the increments Δa, Δb, Δc shall be very small quantities.

A method of successive approximation which makes feasible the reduction of gross errors in the corrections will, I take it, be of considerable interest to mathematical statisticians. Let us, therefore, proceed to the development of a technique which accomplishes precisely this result.

II. **The Theta Technique.** Let us begin our development with the following restatement of the technical problem involved:

Given, the observational series $Y_0, Y_1, \cdots Y_{n-1}$.

Assumed, the function $y = f(x, (a_0 + \theta_1\Delta a), (b_0 + \theta_2\Delta b), (c_0 + \theta_3\Delta c))$.

To find, the values of $\theta_1, \theta_2, \theta_3$.

In this set of relations, a_0, b_0, c_0 and Δa, Δb, Δc are known quantities; while θ_1, θ_2 and θ_3 are each assumed not to exceed ± 1 in value. It follows, therefore,

that the adjusted values of a, b, and c lie within the bounds $a_0 \pm \Delta a$, $b_0 \pm \Delta b$, $c_0 \pm \Delta c$. We may, then, write the following:

$$\begin{aligned} a_1 &= a_0 - \Delta a; \qquad & a_2 &= a_0 + \Delta a. \\ b_1 &= b_0 - \Delta b; & b_2 &= b_0 + \Delta b. \\ c_1 &= c_0 - \Delta c; & c_2 &= c_0 + \Delta c. \end{aligned} \tag{2}$$

The values of θ_1, θ_2 and θ_3 are determined by the following procedure:

First, form the function y from all possible combinations of a_1a_2, b_1b_2, c_1c_2, thus:

$$\begin{aligned} y_{111} &= f(x, a_1b_1c_1). \\ y_{112} &= f(x, a_1b_1c_2). \\ &\cdots\cdots\cdots\cdots \\ y_{222} &= f(x, a_2b_2c_2). \end{aligned} \tag{3}$$

In the case of p parameters, we can evidently form 2^p distinct sets of n values for the function y_{iii}. Since the assigned values of parameters are mere approximations to their true values, each computed set of values for the function y_{iii} will differ from the true values $y = f(x, abc)$.

Second, form the theoretical residuals $y_{iii} - y$, and then compute the corresponding standard errors of estimate σ_{iii}. There will, accordingly, be 2^p values of σ determined, each value being a measure of the error committed in assuming the corresponding approximations to parameters; thus, σ_{111} measures the errors committed in assuming the combination $a_1b_1c_1$; σ_{112} measures the errors committed in assuming $a_1b_1c_2$; $\cdots$; σ_{222} measures the errors committed in assuming $a_2b_2c_2$.

Third, taking the squared reciprocal of σ as a measure of the reliability of a given determination of y_{iii} from the parameters $a_ib_ic_i$, we may form the following comparative tests of the reliability of the 2^p sets of the values of y_{iii}, thus:

$$\begin{aligned} \omega_{111} &= \sigma_{111}^{-2} : (\sigma_{111}^{-2} + \sigma_{112}^{-2} + \cdots + \sigma_{222}^{-2}) = \sigma_{111}^{-2} : \sum \sigma_{iii}^{-2}. \\ \omega_{112} &= \sigma_{112}^{-2} : \sum \sigma_{iii}^{-2}. \\ &\cdots\cdots\cdots\cdots \\ \omega_{222} &= \sigma_{222}^{-2} : \sum \sigma_{iii}^{-2}. \end{aligned} \tag{4}$$

Omega, we shall term the *test constant*. Obviously, $\Sigma\omega_{iii} = 1$.

Fourth, assuming three parameters, let us tabulate the possible subscripts of omega according to the following scheme:

$\omega(a_1)$	$\omega(a_2)$	$\omega(b_1)$	$\omega(b_2)$	$\omega(c_1)$	$\omega(c_2)$
111	211	111	121	111	112
121	221	211	221	211	212
112	212	112	122	121	122
122	222	212	222	221	222

In this table, the subscripts are in the order of abc; so that 111 denotes $\omega(a_1b_1c_1)$; 112 denotes $\omega(a_1b_1c_2)$; etc. Comparing columns $\omega(a_1)$ and $\omega(a_2)$, we observe that the bc subscripts are identical for both; while the a_1 subscripts of the first column are replaced by the a_2 subscripts in the second column. Again, comparing columns $\omega(b_1)$ and $\omega(b_2)$, we see that the ac subscripts are identical for both; while the b_1 subscripts of the one column are replaced by the b_2 subscripts in the other. Finally, comparing columns $\omega(c_1)$ and $\omega(c_2)$, we note that the ab scripts are identical for both; while the c_1 subscripts of the one column are replaced by c_2 subscripts in the other.

Fifth, let us form the column summations $\Sigma\omega(a_1)$, $\Sigma\omega(a_2)$; $\Sigma\omega(b_1)$, $\Sigma\omega(b_2)$; and $\Sigma\omega(c_1)$, $\Sigma\omega(c_2)$. Since the columns $\omega(a_1)$ and $\omega(a_2)$ differ only with respect to the a subscripts, the difference in value between the sums $\Sigma\omega(a_1)$ and $\Sigma\omega(a_2)$ can be due to differences in value between a_1 and a_2 only, and are not at all affected by differences in value between b_1b_2 and c_1c_2. $\Sigma\omega(a_1)$ and $\Sigma\omega(a_2)$ may, therefore, be regarded as the weights of a_1 and a_2 to be used in determining the adjusted value of a; for $\Sigma\omega(a_1) + \Sigma\omega(a_2) = 1$.

We may, then, write the following relations:

$$\begin{aligned} a &= \Sigma\omega(a_1)\cdot a_1 + \Sigma\omega(a_2)\cdot a_2 = \Sigma\omega(a_1)\cdot(a_0 - \Delta a) + \Sigma\omega(a_2)\cdot(a_0 + \Delta a) \\ &= (\Sigma\omega(a_1) + \Sigma\omega(a_2))\cdot a_0 + (\Sigma\omega(a_2) - \Sigma\omega(a_1))\cdot\Delta a = a_0 + \theta(a)\cdot\Delta a. \end{aligned} \tag{5}$$

Since precisely similar reasoning applies to the parameters b_1, b_2 and c_1, c_2, we have the following definitive formulas for computing the values of theta:

$$\begin{aligned} \theta(a) &= \Sigma\omega(a_2) - \Sigma\omega(a_1). \\ \theta(b) &= \Sigma\omega(b_2) - \Sigma\omega(b_1). \\ \theta(c) &= \Sigma\omega(c_2) - \Sigma\omega(c_1). \end{aligned} \tag{6}$$

As the adjusted values of parameters, we have:

$$\begin{aligned} a &= a_0 + \theta(a)\cdot\Delta a. \\ b &= b_0 + \theta(b)\cdot\Delta b. \\ c &= c_0 + \theta(c)\cdot\Delta c. \end{aligned} \tag{7}$$

In this development of the theta technique, we have determined σ_{iii} from the theoretical residuals $y_{iii} - y$. This has served well the purposes of exposition; but, since the true values of the function y are unknown, we must, in practice, compute σ_{iii} from the observational residuals $y_{iii} - Y$. Later in the memoir, it will be shown how the computation of θ may, in numerous cases, be considerably abridged.

Part II

III. **The Principle of Malthus.** Since a determination of the numerical parameters of a given function by means of the theta technique must, at best,

involve a considerable amount of computation, I have chosen for purposes of demonstration a problem which is of much interest in itself. This problem, we shall state in the form of two questions:

First, what is the most appropriate mathematical form of the law of organic growth?

Second, how may the parameters of the indicated function be computed?

Thomas R. Malthus, in his famous essay on *The Principle of Population Growth* assumed that the proportional growth of human populations is properly defined by the differential equation,

$$\frac{1}{p} \cdot \frac{dp}{dt} = b, \tag{8}$$

where p is the population under consideration, t is the measure of time, and b is the stable or geometric rate of growth.

This formula has been destructively criticised on the ground that it fails wholly to give a mathematical description of the manner in which population growth is kept within bounds. So far as any implication of the formula is concerned, populations may grow to infinite magnitudes. An attempt to represent growth by its use must, therefore, result in a succession of discontinuities which are incompatible with the observed facts of organic growth.

IV. **The Symmetric Logistic.** In three memoirs published in 1838, 1845 and 1847, it was suggested by M. Verhulst, Professor of Mathematics in the Ecole Militaire in Brussels, that the rate of population growth might be stated as a function of the population itself. Assuming the limiting value of p to be H, this conception of the growth rate Verhulst expressed by the differential equation,

$$\frac{1}{p} \cdot \frac{dp}{dt} = - b(1 - pH^{-1}). \tag{9}$$

Since this equation expresses proportional growth as a linear function of p, it is the simplest relation of its kind that may be conceived. In representing the rate of growth as a quantity which approaches zero as the population approaches its limiting value, it makes, indeed, a significant advance over the Malthusian formula. Nevertheless, the equation is subject to an interesting limitation, the nature of which is made evident by an examination of the integral form of the function, namely:

$$p = H:[1 + e^{a+bt}]. \tag{10}$$

This we shall now prove to be rotationally symmetric with respect to the point of inflection.

Differentiating equation (9) a second time, we have,

$$\begin{aligned} d^2p &= -b\, dp[p(1 - H^{-1}p)]dt \\ &= p[p^{-2}dp^2 - b\, d^2t + bH^{-1}p\, d^2t + bH^{-1}dp\, dt] \\ &= p^{-1}dp^2 + bH^{-1}p\, dp\, dt. \end{aligned}$$

Hence,

$$\frac{d^2p}{dt^2} = b^2p(1 - H^{-1}p)^2 - b^2H^{-1}p^2(1 - H^{-1}p).$$

Setting $\frac{d^2p}{dt^2} = 0$, we get,

$$1 - 2H^{-1}p = 0.$$

Or

$$p = H/2, \tag{11}$$

which gives the value of p at the point of inflection.

Substituting for p from (10), and solving for t, we have,

$$t_i = -a/b, \tag{12}$$

where t_i is the point of inflection of the function p.

Denoting the magnitude of the population at time t_i by p_i, its magnitude at time t_{i+k} by p_{i+k}, and its magnitude at the time t_{i-k} by p_{i-k}, we have,

$$p_i = H:[1 + e^{a+b(-a/b)}] = H/2. \tag{13}$$

$$p_{i+k} = H:[1 + e^{a+b(t+k\Delta t)}] = H:[1 + e^{bk\Delta t}]. \tag{14}$$

$$p_{i-k} = H:[1 + e^{a+b(t-k\Delta t)}] = H:[1 + e^{-bk\Delta t}]. \tag{15}$$

Measuring p in units of H and setting $u = e^{bk\Delta t}$, we may rewrite these last three equations as follows:

$$H^{-1}p_i = 1/2.$$

$$H^{-1}p_{i+k} = 1:[1 + u].$$

$$H^{-1}p_{i-k} = 1:[1 + u^{-1}].$$

On the hypothesis of rotational symmetry, we have, by subtraction,

$$H^{-1}p_{i+k} - 1/2 = 1/2 - H^{-1}p_{i-k}.$$

In proof, we have:

$$\begin{aligned} 1:[1 + u] &= 1 - 1:[1 + u^{-1}] \\ &= u^{-1}:[1 + u^{-1}] \\ &= 1:[u + 1]. \end{aligned}$$

q. e. d.

Part III

V. **Criticisms of the Logistic.** Because of its symmetric form, many critics have called into question the finality of the logistic as a universal repre-

sentation of population growth. That it applies in particular cases, they contend, is no reason for holding that it must apply in general. Professors Raymond Pearl and Lowell J. Reed of Johns Hopkins University—to whom we are indebted for the rediscovery of the earlier researches of Verhulst—have proposed, as the proper form of the generalized growth curve, the following function:

$$p = H:[1 + e^{a+bt+ct^2+dt^3}]. \tag{16}$$

In their view, this equation is suited not only to representing a single cycle of growth, but two successive cycles as well. This claim, however, must be rejected; for, if true, it would mean that one cycle of growth is predictable from another, a circumstance which is clearly inconsistent with the assumptions laid down by these same investigators.

Moreover, so far as I can learn from their published writings, these authors have never considered the implications of the differential form of the function they propose.

Differentiating (16), we have,

$$\frac{1}{p} \cdot \frac{dp}{dt} = -\ (b + 2cx + 3dx^2)\ (1 - H^{-1}p).$$

Here, we find the stable growth constant of Malthus replaced by an expression which is quadratic in t. This means that, for a population which is freed of a restraining limit, proportional growth tends generally toward infinite values. If there are any facts to support such a conception of organic growth, I do not know what they are, and must, perforce, reject the contention that equation (16) is the generalized form of the Verhulst function.

VI. **Fundamental Assumptions.** In order to represent the phenomenon of population growth mathematically, I hold the following assumptions to be necessary:

(a) Under favoring conditions, population may increase at a constant geometric rate.

(b) Under all circumstances, the rate of growth must be a finite and continuous quantity.

(c) The magnitude of a population is always a positive, real number.

(d) The growth of population tends toward restriction within definite bounds.

(e) The growth of population is a function of time.

(f) The basic conditions of growth are free of cataclysmic disturbances.

The first of these assumptions is given in recognition of well known facts concerning organic growth. The second is necessary because, even when the size of a population is freed of definite restriction, the pattern of growth is not necessarily geometric. The third assumption affirms the absurdity of representing a population as a negative or infinite quantity. The fourth merely asserts the indisputable fact that the organism must always grow in a finite environment. The fifth gives place to the concept of growth as the resultant of a complex of

causes, no one of which can be isolated as an entirely independent variable. While the final assumption recognizes that major disturbing influences may profoundly affect the course of growth.

VII. **The Skew Logistic.** In accord with our fundamental assumptions, we may form the following differential equations:

$$\frac{1}{p'} \cdot \frac{dp'}{dt} = -\,[b + sm \cdot \cos\,(m(t+q))]\,[1 - H^{-1}p'] \qquad \text{Type } \alpha$$

$$= -\,[b + 2sm^2(t+q) : (1 + m^2(t+q)^2)]\,[1 - H^{-1}p'] \qquad \text{Type } \beta \quad (17)$$

$$= -\,[b + sm^2(t+q) : \sqrt{1 + m^2(t+q)^2}]\,[1 - H^{-1}p'] \qquad \text{Type } \gamma$$

In these equations, $p' = p - L$, and measures p from its lower limit as origin. On separating variables, the following integrations may be performed:

$$-\int [dp' : (p'(1 - H^{-1}p'))] = -\log\,[p' : (1 - H^{-1}p')] = \log\,[(H - p') : (Hp')].$$

Writing $z = m(t+q)$, $dz = mdt$; so that we have:

$$b\int dt + s\int \cos z\, dz = A + bt + s \cdot \sin z.$$

$$b\int dt + 2s\int [z : (1 + z^2)]dz = A + bt + s \cdot \log\,(1 + z^2).$$

$$b\int dt + s\int [z : \sqrt{1 + z^2}]dz = A + bt + s\sqrt{1 + z^2}.$$

From these integrals, we form the following equations:

$$\log\,[(H - p') : (Hp')] = A + bt + s \cdot \sin\,[m(t+q)].$$

$$\log\,[(H - p') : (Hp')] = A + bt + s \cdot \log\,[1 + m^2(t+q)^2].$$

$$\log\,[(H - p') : (Hp')] = A + bt + s \cdot \sqrt{1 + m^2(t+q)^2}.$$

We have, finally, on taking antilogarithms and making the substitutions $p = p' + L$, $a = A - \log H$:

$$p = L + H : [1 + e^{a + bt + s \cdot \sin(m(t+q))}]. \qquad \text{Type } \alpha$$

$$p = L + H : [1 + e^{a + bt + s \cdot \log(1 + m^2(t+q)^2)}]. \qquad \text{Type } \beta \quad (18)$$

$$p = L + H : [1 + e^{a + bt + s\sqrt{1 + m^2(t+q)^2}}]. \qquad \text{Type } \gamma$$

These equations give the normal forms of the skew logistic.

VIII. **Properties of the Skew Logistic.** We may deduce the properties of the skew logistic by examining both its differential and integral forms. Considering the derivative of Type α, we note that the Malthusian constant b is

replaced by a trigonometric function whose amplitude is $b \pm sm$, and whose phase depends on the values of m and q. When $b \pm sm = 0$, the derivative must also equal zero, and a flat point in the curve of p is indicated. When b is absolutely less than sm, the derivative changes sign and the curve of p reverses its direction. Thus, the integral form of Type α modifies the symmetric form of the logistic by a succession of minor cycles in which the rate of growth is alternately accelerated and retarded.

Considering Type β, we find the Malthusian constant replaced by a function whose maximum and minimum values are attained when $t = m^{-1} - q$. Obviously, therefore, this function passes through a single period whose amplitude is $b \pm sm$, and whose phases are b, $b + sm$, b, $b - sm$, b. When $b \pm sm = 0$, a flat point in the curve of p is generated. The effect of skewness on the rate of growth passes through two double phases. Where b and s are of the same sign, these phases are: first, increasing retardation followed by decreasing retardation when $t + q$ is negative; and, second, increasing acceleration followed by decreasing acceleration when $t + q$ is positive. Where b and s are of opposite sign, the corresponding phases are: first, increasing acceleration followed by decreasing acceleration when $t + q$ is negative; and, second, increasing retardation followed by decreasing retardation when $t + q$ is positive. It is to be noted that, when sm is absolutely greater than b, the derivative will change sign twice before the upper limit is reached. Under these circumstances, the function p passes through a double reversal of direction.

Considering Type γ, we find the Malthusian constant of the derivative replaced by a function which is aperiodic and which approaches the limits $b \pm sm$ as t approaches $\pm \infty$. When b and s are of the same sign, skewness passes through the two following phases: first, the phase of decreasing retardation when $t + q$ is negative; and, second, the phase of increasing acceleration when $t + q$ is positive. On the other hand, when b and s are of opposite sign, the corresponding phases are: first, that of decreasing acceleration when $t + q$ is negative; and, second, that of increasing retardation when $t + q$ is positive. When sm is absolutely greater than b, the derivative changes sign, and the function p passes from a continuously increasing phase to a continuously decreasing phase, or *vice versa*.

In general, it may be said of all three types—α, β and γ—that, if the derivative is not restricted to a single change of sign, L denotes a lower asymptote of the function p; while, under the same conditions, H denotes the higher limit approached by the function $p - L$. When H is negative, the effect is to make L an upper, and $L - H$ a lower, asymptote of the curve p.

In the case of Type γ, when the function p makes a single change of sign, either H or L becomes a maximum (or minimum) value instead of an asymptote of the curve. In this event, it will be noted that the factor $1 - H^{-1}p$ appearing in the derivative does not approach zero as a limit with increasing values of t, but rather passes through a minimum and then approaches the limit 1 in either direction.

The parameter s may be positive or negative in sign, and is termed the index of skewness or, briefly, the *skewness* of the function. Obviously, m is always positive, and, since it determines the rate at which skewness develops, is properly termed the *development*. The point in time at which skewness passes from an accelerating to a retarding phase, or *vice versa*, is fixed by the value of q, which is, therefore, termed the *transition*. The parameter b, as has already been stated, is termed the stable growth tendency or, technically, the *stability* of the function. And since the position of the curve p on an arbitrary time scale will vary with the value of a, this parameter I have designated the *location*.

In all three types of the skew logistic, if $e^{\psi(t)}$ is a continuously decreasing function and both H and L are positive, the curve of p may be described as of the *rising hillside form*. In the case of Type γ, if the derivative changes from positive to negative sign, the curve may be described as *mountain formed*. If $e^{\psi(t)}$ increases continuously, the curve is of the *falling hillside* variety, except when the derivative of Type γ changes from negative to positive sign, in which event a *valley form* is generated.

Part IV

IX. **Parameters of the Symmetric Logistic.** The numerical parameters of the symmetric logistic (10) are most easily determined by the method of differences. First, we write,

$$p_i^{-1} = C + e^{A+bt}, \tag{19}$$

where $C = H^{-1}$; $A = a - \log H$; and $i = 0, 1, 2, \cdots n - 1$.

Assuming Δt constant, let us give to t the increment $k\Delta t$, thus:

$$p_{i+k}^{-1} = C + e^{A+b(t+k\Delta t)}. \tag{20}$$

Subtracting (19) from (20), we obtain

$$\Delta_k p_i^{-1} = e^{A+b(t+k\Delta t)} - e^{A+bt} = Be^{A+bt}, \tag{21}$$

where $B = e^{bk\Delta t} - 1$. The quantity $\Delta_k p_i^{-1} = p_{i+k}^{-1} - p_i^{-1}$ is termed a first order difference of rank k.

Giving to t in equation (21) the increment $k\Delta t$, we get

$$\Delta_k p_{i+k}^{-1} = Be^{A+b(t+k\Delta t)}. \tag{22}$$

Dividing (22) by (21), we have,

$$\Delta_k p_{i+k}^{-1} : \Delta_k p_i^{-1} = e^{bk\Delta t}.$$

Taking logarithms, we obtain

$$\Delta_k \log \Delta_k p_i^{-1} = \log \Delta_k p_{i+k}^{-1} - \log \Delta_k p_i^{-1} = bk\Delta t,$$

which defines the parameter b. We can form $n - 2k$ such equations. Hence,

b is uniquely determined by the relation

$$b = [\sum_{i=0}^{i=n-2k-1} \Delta_k \log \Delta_k P_i^{-1}]:[k(n - 2k)\Delta t]$$
$$= [\sum_{i=k}^{i=n-k-1} \log \Delta_k P_i^{-1} - \sum_{i=0}^{i=n-2k-1} \log \Delta_k P_i^{-1}]:[k(n + 2k)\Delta t], \quad (23)$$

where $k = n:3$ to the nearest integer.

Returning to (21), we have the following relation determining the value of A:

$$A = \log [\sum_{i=0}^{i=n-k-1} \Delta_k P_i^{-1}] - \log [B \sum_{i=0}^{i=n-k-1} e^{bt}]$$
$$= \log [\sum_{i=k}^{i=n-1} P_i^{-1} - \sum_{i=0}^{i=n-k-1} P_i^{-1}] - \log [B \sum_{i=0}^{i=n-k-1} e^{bt}], \quad (24)$$

where $k = n:2$ to the nearest integer.

From equation (19), we have

$$C = [\sum_{i=0}^{i=n-1} P_i^{-1} - e^A \sum_{i=0}^{i=n-1} e^{bt}]:n. \quad (25)$$

The values of H and a are, obviously, given by

$$H = C^{-1}. \quad (26)$$

$$a = A + \log H. \quad (27)$$

In the relations defining b, A and C, the values of P must be obtained from the observations. In computing the values of k, the formula is:

$$k = n(r + 1)^{-1},$$

where n is the number of observations, and r denotes the order of reduction involved in the defining relation.

In my first treatment of the subject, I assumed that the value of k for all orders of reduction might be determined from the reduction of highest order involved; but I have since found that I erred in this view. The point is that the function $\psi(p) = k^r(n - rk)$, discussed in the original memoir, must be maximized with respect to k separately for each order of difference involved; or, in other words, the rank constant k must be given a separate determination for each parameter defined if the most accurate results are to be obtained.

X. **Parameters of the Skew Logistic.** I shall now show how the method of differences may be used to abridge the computations involved in applying the theta technique to the determination of the parameters of the skew logistic. In this, as in the preceding section, we assume Δt constant.

Operating on Type γ of equation (18), we write

$$p_i = L + H:[1 + e^{a+bt+s\sqrt{1+m^2(t+q)^2}}]. \quad (28)$$

To begin with, let us write the transformation of ordinate

$$G = \log [H(p - L)^{-1} - 1].$$

Also, let us write

$$F = \sqrt{1 + m^2(t + q)^2}.$$

We may now rewrite equation (28) in the form

$$G_i = a + bt + sF_i. \tag{29}$$

Giving to t the increment $k\Delta t$, we have

$$G_{i+k} = a + b(t + k\Delta t) + sF_{i+k}. \tag{30}$$

Subtracting (29) from (30), we have,

$$\Delta_k G_i = bk\Delta t + s\Delta_k F_i. \tag{31}$$

Again giving to t the increment $k\Delta t$, we obtain

$$\Delta_k G_{i+k} = bk\Delta(t + k\Delta t) + s\Delta_k F_{i+k}. \tag{32}$$

Subtracting (31) from (32), we obtain

$$\Delta_k G_{i+k} - \Delta_k G_i = (bk\Delta t - bk\Delta t) + s(\Delta_k F_{i+k} - \Delta_k F_i),$$

or

$$\Delta_k^2 G_i = s\Delta_k^2 F_i. \tag{33}$$

We can form $n - 2k$ such equations, and may, therefore, form $n - 2k$ approximations to the value of the parameter s, as follows:

$$s_i = [\Delta_k^2 G_i]:[\Delta_k^2 F_i]; \qquad i = 0, 1, \cdots, n - 2k - 1.$$

Taking the mean value of the set s_i as its most probable value, we have,

$$s_0(HL\cdot mq) = \Sigma s_i:(n - 2k);\ k = n:3 \text{ to the nearest integer} \tag{34}$$

In this determination of s_0, the only parameters directly involved are H, L, m and q, the parameters a and b having been eliminated. By assigning values to H_0, L_0, m_0 and q_0, we may, on setting up the arbitrary corrections ΔH, ΔL, Δm and Δq, write down the following:

$$\begin{aligned} &H_1 = H_0 - \Delta H; && H_2 = H_0 + \Delta H; && L_1 = L_0 - \Delta L; && L_2 = L_0 + \Delta L; \\ &m_1 = m_0 - \Delta m; && m_2 = m_0 + \Delta m; && q_1 = q_0 - \Delta q; && q_2 = q_0 + \Delta q. \end{aligned}$$

Since s_0 is a function of H, L, m and q, we may, by entering the subscripts of the combination $HL\cdot mq$, tabulate the possible determinations of s_0 as follows:

11·11	11·12	11·21	11·22
12·11	12·12	12·21	12·22
21·11	21·12	21·21	21·22
22·11	22·12	22·21	22·22

In this tabulation, the subscripts of parameters are in the order of $HL\cdot mq$; so that 12·21 denotes $s_0(H_1L_2\cdot m_2q_1)$, etc.

From the table, it is seen that we may compute $2^4 = 16$ distinct sets of approximations to $s_0(HL\cdot mq)$. Since the true values of H, L, m and q are unknown, each set of approximations s_i will show a characteristic variation about its mean

value, s_0. This variation is most conveniently measured by the mean deviation

$$\epsilon = (s_0 - s_0')2N':N = (s_0'' - s_0)2N'':N, \tag{35}$$

where the second relation serves as a check on the computation by the first; $N = n - 2k$; N' denotes the number of items s_i which are *less* than s_0 in value, and N'', the number of items s_i which are *greater* than s_0 in value; while s_0' denotes the mean of the N' values of s_i which are less than s_0, and s_0'', the mean of the N'' values of s_i which are greater than s_0.

The reliability of a given value of s_0 as a measure of the central tendency of the corresponding set s_i is sufficiently determined by ϵ^{-2}, which serves at the same time to measure the reliability of the combination $HLmq$ figuring in the computation of the given set s_i. We may, therefore, compute the values of the test constant, ω, directly from the values of ϵ^{-2} by means of the relation,

$$\omega(HL\cdot mq) = \epsilon_j^{-2}:[\epsilon_{11\cdot 11}^{-2} + \epsilon_{11\cdot 12}^{-2} + \cdots + \epsilon_{22\cdot 22}^{-2}] = \epsilon_j^{-2}:\Sigma\epsilon^{-2}, \tag{36}$$

where $j = 11\cdot 11, 11\cdot 12, \cdots, 22\cdot 22; \Sigma\omega = 1$.

Since four values of theta are to be determined, we must arrange the sixteen values of omega in four ways, as shown by the following tabulation of subscripts:

$\omega(H_1)$	$\omega(H_2)$	$\omega(L_1)$	$\omega(L_2)$	$\omega(m_1)$	$\omega(m_2)$	$\omega(q_1)$	$\omega(q_2)$
11·11	21·11	11·11	12·11	11·11	11·21	11·11	11·12
11·12	21·12	11·12	12·12	11·12	11·22	11·21	11·22
11·21	21·21	11·21	12·21	12·11	12·21	12·11	12·12
11·22	21·22	11·22	12·22	12·12	12·22	12·21	12·22
12·11	22·11	21·11	22·11	21·11	21·21	21·11	21·12
12·12	22·12	21·12	22·12	21·12	21·22	21·21	21·22
12·21	22·21	21·21	22·21	22·11	22·21	22·11	22·12
12·22	22·22	21·22	22·22	22·12	22·22	22·21	22·22

Knowing the values of omega, we have at once,

$$\theta(H) = \Sigma\omega(H_2) - \Sigma\omega(H_1); \quad \theta(L) = \Sigma\omega(L_2) - \Sigma\omega(L_1);$$

$$\theta(m) = \Sigma\omega(m_2) - \Sigma\omega(m_1); \quad \theta(q) = \Sigma\omega(q_2) - \Sigma\omega(q_1). \tag{37}$$

$$H = H_0 + \theta(H)\cdot\Delta H; \quad L = L_0 + \theta(L)\cdot\Delta L;$$

$$m = m_0 + \theta(m)\cdot\Delta m; \quad q = q_0 + \theta(q)\cdot\Delta q. \tag{38}$$

The process of adjustment should be repeated until errors in the parameters diminish to negligible proportions.

With H, L, m and q known to a sufficient approximation, we may form anew the functions $G(H, L, m, q)$ and $F(H, L, m, q)$. We can then write $n - 2k$ equations of form (33), viz.:

$$\Delta_k^2 G_i = s\Delta_k^2 F_i.$$

Summing these equations, we have,

$$\Sigma\Delta_k^2 G_i = s\Sigma\Delta_k^2 F_i, \tag{39}$$

where $\Sigma\Delta_k^2 G_i = \sum_{i=2k}^{i=n-1} G_i - 2\sum_{i=k}^{i=n-k-1} G_i + \sum_{i=0}^{i=n-2k-1} G_i$;

$$\Sigma\Delta_k^2 F_i = \sum_{i=2k}^{i=n-1} F_i - 2\sum_{i=k}^{i=n-k-1} F_i + \sum_{i=0}^{i=n-2k-1} F_i\,;$$

where $k = n{:}3$ to the nearest integer.

The approximate value of s is now obtained from the relation

$$s = [\sum_{i=0}^{i=n-2k-1} \Delta_k^2 G_i]:[\sum_{i=0}^{i=n-2k-1} \Delta_k^2 F_i]. \tag{40}$$

Returning to equation (31), we solve for $bk\Delta t$, obtaining,

$$bk\Delta t = \Delta_k G_i - s\Delta_k F_i\,.$$

Since we can form $n - k$ such equations, the approximate value of b is given by the relation

$$b = [\Sigma\Delta_k G_i - s\Sigma\Delta_k F_i]:[k(n-k)\Delta t] \tag{41}$$

$$= [(\sum_{i=k}^{i=n-1} G_i - \sum_{i=0}^{i=n-k-1} G_i) - s(\sum_{i=k}^{i=n-1} F_i - \sum_{i=0}^{i=n-k-1} F_i)]:[k(n-k)\Delta t],$$

where $k = n{:}2$ to the nearest integer.

From equation (29), we obtain the approximate value of a as follows:

$$a = [\sum_{i=0}^{i=n-1} G_i - b\sum_{i=0}^{i=n-1} t - s\sum_{i=0}^{i=n-1} F_i]:n. \tag{42}$$

Comparing the abridged method of computing the values of theta here outlined with the general procedure of section II, it will be seen that we have been able to reduce the number of values of omega which it is necessary to determine from $2^7 = 128$ to $2^4 = 16$. In cases where L may be assumed to equal zero, the number of values of omega which must be computed is further reduced to $2^3 = 8$.

Part V

XI. **Symmetric Parameters for the Population of the United States.** I have determined the numerical values of the parameters of both the symmetric and the skew forms of the logistic from the population figures for the United States given by the Bureau of the Census. The only departure in the data from the census figures consists in the interpolation of all items to June 1st as the date of observation. The values of the symmetric parameters are computed from the data of Table I, as follows.

Setting $k = 15 \div 3 = 5$, we have, by equation (23),

$$\sum_0^4 \Delta_5 \log \Delta_5 P_i^{-1} = \sum_5^9 \log \Delta_5 P_i^{-1} - \sum_0^4 \log \Delta_5 P_i^{-1}$$

$$= \bar{9}.71878n - \bar{5}.14555n = -3.42677.$$

TABLE I

Data for the Symmetric Logistic

i	P^{-1}	$\Delta_5 P^{-1}$	$\log \Delta_5 P^{-1}$	10^{bt}
0	0.25582	−0.19724	$\bar{1}.29500_n$	1.00000
1	0.18939	−0.14627	$\bar{1}.16516_n$	0.72934
2	0.13885	−0.10704	$\bar{1}.02955_n$	0.53193
3	0.10431	−0.07838	$\bar{2}.89421_n$	0.38796
4	0.07770	−0.05776	$\bar{2}.76163_n$	0.28295
5	0.05858	−0.04269	$\bar{2}.63033_n$	0.20637
6	0.04312	−0.02996	$\bar{2}.47654_n$	0.15051
7	0.03181	−0.02098	$\bar{2}.32181_n$	0.10978
8	0.02593	−0.01650	$\bar{2}.21748_n$	0.08006
9	0.01994	−0.01182	$\bar{2}.07262_n$	0.05839
10	0.01589			0.04259
11	0.01316			0.03106
12	0.01083			0.02265
13	0.00943			0.01652
14	0.00812			0.01205
Σ	1.00288	−0.70864	$\overline{14}.86433_n$	3.66216

TABLE II(A)

Data for the Skew Logistic

i	G_1	G_2	F_{11}	F_{12}	F_{21}	F_{22}
0	+ 1.67998	+ 1.71132	6.47765	3.35261	9.65194	4.90306
1	+ 1.54968	+ 1.57779	5.68859	2.60000	8.45931	3.73631
2	+ 1.40690	+ 1.43878	4.90306	1.88680	7.26911	2.60000
3	+ 1.27698	+ 1.30927	4.12311	1.28062	6.08276	1.56205
4	+ 1.14130	+ 1.17416	3.35261	1.00000	4.90306	1.00000
5	+ 1.00816	+ 1.04179	2.60000	1.28062	3.73631	1.56205
6	+ 0.85948	+ 0.89428	1.88680	1.88680	2.60000	2.60000
7	+ 0.70540	+ 0.74193	1.28062	2.60000	1.56205	3.73631
8	+ 0.59699	+ 0.63515	1.00000	3.35261	1.00000	4.90306
9	+ 0.44841	+ 0.48956	1.28062	4.12311	1.56205	6.08276
10	+ 0.30840	+ 0.35346	1.88680	4.90306	2.60000	7.26911
11	+ 0.17992	+ 0.22981	2.60000	5.68859	3.73631	8.45931
12	+ 0.02885	+ 0.08647	3.35261	6.47765	4.90306	9.65194
13	− 0.09590	− 0.02968	4.12311	7.26911	6.08276	10.84620
14	− 0.25808	− 0.17670	4.90306	8.06226	7.26911	12.04159
Σ	+10.83647	+11.47739	49.45864	55.76384	71.41783	80.95375

TABLE II(B)

Data for the Skew Logistic

i	$\Delta_5^2 G_1$	$\Delta_5^2 G_2$	$\Delta_5^2 F_{11}$	$\Delta_5^2 F_{12}$	$\Delta_5^2 F_{21}$	$\Delta_5^2 F_{22}$
0	−0.02794	−0.01880	3.16445	5.69443	4.77932	9.04807
1	+0.01064	+0.01904	4.51499	4.51499	6.99562	6.99562
2	+0.02495	+0.04139	5.69443	3.16445	9.04807	4.77932
3	−0.01290	+0.00929	6.24622	1.84451	10.16552	2.60213
4	−0.01360	+0.01834	5.69443	0.81604	9.04807	0.87607
Σ	−0.01885	+0.06926	25.31452	16.03442	40.03660	24.30121

We note that $k(n - 2k)\Delta t = 5(15\text{–}10)1 = 25$; hence,

$$b = -3.42677 \div 25 = -0.1370708.$$

Next, set $k = 15 \div 2 = 7$, to the nearest integer; then, by equation (24), we get

$$\textstyle\sum_0^7 \Delta_7 P_i^{-1} = \sum_7^{14} P_i^{-1} - \sum_0^7 P_i^{-1} = 0.10330 - 0.86777 = -0.76447;$$

$$B = 10^{bk\Delta t} - 1 = 10^{-0.1370708\times 7} - 1 = -0.89022; \qquad \textstyle\sum_0^7 10^{bt} = 3.39884.$$

Hence,

$$A = \log[-0.76447] - \log[-0.89022 \times 3.39884] = \bar{1}.4025324.$$

We have next

$$\textstyle\sum_0^{14} P_i^{-1} = 1.00288; \qquad \sum_0^{14} 10^{bt} = 3.66216; \qquad 10^A = 0.25266.$$

By equation (25), then, we obtain

$$C = [1.00288 - 0.25266 \times 3.66216] \div 15 = 0.0051747.$$

By equation (26), we get

$$H = C^{-1} = 193.25.$$

Finally, by equation (27), we obtain

$$a = A + \log H = \bar{1}.4025324 + 2.2861136 = 1.68865.$$

The point of inflection of the curve is given by

$$t_i = -a{:}b = 1.68865 \div 0.1370708 = 12.319.$$

XII. **Skew Parameters for the Population of the United States.** Assuming $L = 0$, we form

$$H_1 = 198.0 - 7.0 = 191.0; \qquad H_2 = 198.0 + 7.0 = 205.0.$$

$$m_1 = 1.0 - 0.2 = 0.8; \qquad m_2 = 1.0 + 0.2 = 1.2.$$

$$q_1 = -6.0 - 2.0 = -8.0; \qquad q_2 = -6.0 + 2.0 = -4.0.$$

Next, the primary data of Tables II(a) and II(b) are computed. Setting $k = 15 \div 5 = 3$, $n - k$ values of the 2^3 sets of s_i are determined and entered in Table III(a). The values of s_0, ϵ and ω for each set are computed by equations (34), (35) and (36).

In Table III(b), the several values of ω are arranged according to their association: first, with H_1, H_2; second, with m_1, m_2; and, third, with q_1, q_2. The column sums yield the weights $\Sigma\omega$. The values of θ and the adjusted values of parameters are computed by equations (37) and (38):

TABLE III(A)

Data for the Computation of Θ

i	$s(1.11)$	$s(1.12)$	$s(1.21)$	$s(1.22)$	$s(2.11)$	$s(2.12)$	$s(2.21)$	$s(2.22)$
0	−0.00883	−0.00491	−0.00585	−0.00309	−0.00594	−0.00330	−0.00393	−0.00208
1	+0.00236	+0.00236	+0.00152	+0.00152	+0.00422	+0.00422	+0.00272	+0.00272
2	+0.00438	+0.00788	+0.00276	+0.00522	+0.00727	+0.01308	+0.00457	+0.00866
3	−0.00207	−0.00699	−0.00127	−0.00496	+0.00149	+0.00504	+0.00091	+0.00357
4	−0.00239	−0.01667	−0.00150	−0.01552	+0.00322	+0.02247	+0.00203	+0.02093
Σ	−0.00655	−0.01832	−0.00434	−0.01683	+0.01026	+0.04151	+0.00630	+0.03380
s_0	−0.00131	−0.00366	−0.00087	−0.00337	+0.00205	+0.00830	+0.00126	+0.00676
ϵ	+0.00374	+0.00703	+0.00241	+0.00550	+0.00342	+0.00404	+0.00222	+0.00643
ω	+0.10624	+0.03012	+0.25711	+0.04919	+0.12708	+0.09137	+0.00290	+0.03061

TABLE III(B)

Data for the Computation of Θ

	$\omega(h_1)$	$\omega(h_2)$	$\omega(m_1)$	$\omega(m_2)$	$\omega(q_1)$	$\omega(q_2)$
	0.1062	0.1271	0.1062	0.2571	0.1062	0.0301
	0.0301	0.0914	0.0301	0.0492	0.2571	0.0492
	0.2571	0.3029	0.1271	0.3029	0.1271	0.0914
	0.0492	0.0360	0.0914	0.0360	0.3029	0.0360
Σ	0.4426	0.5574	0.3548	0.6452	0.7933	0.2067

TABLE IV(A)

Summary of Adjustments

Parameter	Estimated Value	Δ	Θ	$\Delta\cdot\Theta$	Adjusted Value
H	+198.0	+7.0	+0.1148	+0.8036	+198.80
m	+ 1.0	+0.2	+0.2904	+0.05808	+1.05808
q	− 6.0	+2.0	−0.5866	−1.1732	−7.1732

TABLE IV(B)

Final Transformations

i	$G(Hmq)$	$F(Hmq)$
0	1.69772	7.65559
1	1.56410	6.60800
2	1.42495	5.46440
3	1.29526	4.52752
4	1.15991	3.50336
5	1.02722	2.50753
6	0.87921	1.59408
7	0.72613	1.01784
8	0.61866	1.32865
9	0.47182	2.17626
10	0.33408	3.15374
11	0.20842	4.17075
12	0.06189	5.20418
13	$\bar{1}$.94223	6.24580
14	$\bar{1}$.78913	7.29229
Σ	11.20073	62.54999

Finally, the functions G and F are formed anew from the adjusted values of H, m, q. The adjusted values of s, b and a are computed by equations (40), (41) and (42), as follows:

$$s = [\textstyle\sum_{10}^{14} G_i - 2\sum_5^9 G_i + \sum_0^4 G_i] : [\sum_{10}^{14} F_i - 2\sum_5^9 F_i + \sum_0^4 F_i]$$
$$= [0.33574 - 2 \times 3.72304 + 7.14194] \div [26.06676 - 2 \times 8.62436 + 27.85887]$$
$$= 0.03161 \div 36.67691 = 0.00086185.$$
$$b = [\textstyle\sum_8^{14} G_i - \sum_0^6 G_i - s(\sum_8^{14} F_i - \sum_0^6 F_i)] : [k(n - k)\Delta t]$$
$$= [1.42623 - 9.04837 - 0.00086185(29.57167 - 31.96048)] \div [7(15 - 7)1]$$
$$= [-7.62214 - 0.00086185 \times (-2.38881)] \div [56] = -0.13607.$$
$$a = [\textstyle\sum_0^{14} G_i - b\sum_0^{14} t - s\sum_0^{14} F_i] : n$$
$$= [11.20073 - (-0.13607 \times 105) - 0.00086185 \times 62.54999] \div 15 = 1.69561.$$

In the present case, the values of H_0, m_0 and q_0 were known within definite limits from previous experimentation. The values of the corrections, $\theta \cdot \Delta$, were, on this account, smaller than should ordinarily be expected from a first application of the technique. Always, it is necessary to take Δ sufficiently large to insure $\theta < 1$. As a preliminary step, it is not infrequently advantageous to compute trial values of ϵ by holding constant each two of the parameters H_0, m_0 and q_0 while experimenting roughly with the third.

TABLE V(A)

Ordinates of Fitted Curves

Year	Census Count	Symmetric Ordinates	Percentage Deviations	Skew Ordinates	Percentage Deviations
1790	3.909	3.88	−0.78	3.87	−0.01
1800	5.280	5.28	−0.03	5.27	−0.25
1810	7.202	7.16	−0.52	7.15	−0.73
1820	9.587	9.69	+1.07	9.67	+0.88
1830	12.866	13.04	+1.37	13.02	+1.20
1840	17.069	17.45	+2.22	17.42	+2.09
1850	23.192	23.15	−0.20	23.13	−0.28
1860	31.443	30.38	−3.36	30.37	−3.42
1870	38.558	39.36	+2.09	39.31	+1.95
1880	50.156	50.18	+0.05	50.07	−0.18
1890	62.948	62.61	−0.31	62.60	−0.55
1900	75.995	76.79	+1.05	76.64	+0.86
1910	92.329	91.76	−0.62	91.72	−0.67
1920	106.001	106.96	+0.90	107.16	+1.09
1930	123.068	121.66	−1.14	122.23	−0.69

TABLE V(B)

Extrapolations

Year	Forecast	Sym. O.	Sk. O.	Year	Sym. O.	Sk. O.
1940	137.20	135.22	136.26	1780	2.844	2.850
1950	149.29	147.18	148.78	1770	2.083	2.095
1960	159.88	157.33	159.52	1760	1.523	1.539
1970	168.71	165.66	168.42	1750	1.113	1.130
1980	175.83	172.33	175.59	1740	0.813	0.829
1990	181.46	177.52	181.25	1730	0.594	0.608
2000	185.82	181.52	185.63	1720	0.434	0.445
2010	189.14	184.55	188.98	1710	0.316	0.280
2020	193.11	186.82	192.97	1700	0.231	0.238
2030	193.54	188.52	193.40	1690	0.168	0.173
2040	194.94	189.77	194.83	1680	0.123	0.127
2050	195.98	190.72	195.87	1620	0.090	0.092
2060	196.75	191.39	196.64	1610	0.065	0.067
2070	197.31	191.88	197.22	1600	0.048	0.049
2080	197.73	192.25	197.64	1590	0.035	0.036
2090	198.03	192.52	197.94			
2100	198.25		198.17			
2110	198.42		198.34			
2120	198.54		198.46			
2130	198.63		198.55			

Part VI

XIII. **General Considerations.** The technique of solution for the numerical values of parameters presented in the foregoing pages is generally applicable to continuous functions of real variables. The abridged procedure may be followed whenever the given function involves a component which is linear in certain of the parameters: for, in such cases, it is always possible to effect a transformation of ordinates which will permit of the elimination of the parameters of the linear component. In any event, the equation of the function may be solved for a single parameter which may then be employed, as in our illustration, as a means of determining the values of the test constant, omega.

XIV. **An Interpretation of Results.** The equations of the symmetric and skew logistic curves as computed for the population of the United States are, written to the natural base, as follows:

$$p = 193.25:[1 + e^{3.88826-0.31562t}].$$

$$p = 198.80:[1 + e^{3.90429-0.31331t+0.0019845\sqrt{1+1.0581^2(t-7.1732)^2}}]$$

The amount of skewness in the second of these equations, as measured by the value of s, is small; but, owing to the fair size of the parameter m, it develops rapidly and affects the form of the curve sensibly. The major effect is to raise the value of the limiting population as given in the first equation by about six millions and to prolong the period of growth by about forty years. The approximate limit of 193 millions in the symmetric form is reached about the year 2090; while the approximate limit of 199 millions of the skew form is not arrived at until about the year 2130.

The positive sign of s makes for a decreasing acceleration of the rate of increase during the earlier phases of growth and for an increasing retardation of this rate during the later phases, the value of q fixing the point of transition in the year 1861. This general epoch has often been cited by sociologists as marking the shift from a dominantly rural-agricultural civilization to a dominantly urban-industrial one. The point at which the change takes place has, to my knowledge, never before been defined mathematically.

Both curves fit the observations excellently, as shown by the percentage deviations of Table V(a). The forecasted growth presented in Table V(b) is based on the skew ordinates, the formula being

$$P_t = p_t(P_{14}/p_{14})^{1/(t-14)}, \tag{43}$$

where P denotes the actual population series, observed or predicted, and p, the skew ordinates. The assumptions of the formula are two: first, that it is the observed population P_{14} which initiates the forecasted series; and, second, that the influence of the correction factor P_{14}/p_{14} diminishes with the time.

The extrapolations of both the skew and symmetric formulas contrast with the results obtained by Doctors Dublin and Lotka, who predict a stationary

population of 150 millions by 1970. For the same year, the ordinates of both the skew and symmetric curves exceed this figure, the one by 15.66, and the other by 18.42 millions.

The limit of 150 millions referred to was arrived at by analysis of current tendencies in birth and death rates. The argument is that current birth rates are spuriously high and current death rates spuriously low because of the abnormally high proportion of men and women in the reproductive ages. This circumstance is due, in part, to the influx in the past of immigrants from communities having a high normal birth rate, and, in part, to the high birth rates of preceding generations of parents in this country.

After computation of the necessary corrections has been made, the true rate of natural increase of the white population for the registration area of the United States for the year 1920 is seen to be only about 5.4 per thousand instead of the 10.7 per thousand indicated by the crude rates. For the year 1930, the actual rate of increase is 7.5 per thousand; while the corrected or true rate turns out to be virtually zero. Under the interpretation of the authorities cited, the spurious excess of births over deaths will be entirely dissipated by the year 1970, with the result of the stationary population predicted.

The hazard peculiar to this method of inference arises from two assumptions that are made: first, that the present collection and registration of vital statistics is sufficiently reliable to make precise estimate of the true rate of natural increase possible; second, that the tendencies of fecundity and mortality exhibited by current data are stable.

With respect to the first assumption, the authors have this to say:

"One factor of safety of unknown magnitude remains. There is still some degree of laxity in the registration of births, and the figures of the true rate of natural increase may, on that account, be somewhat larger than recorded above."

The caution of the authors in this statement is in contrast with the uncritical acceptance of their results by those who fail to grasp the implications of technique.

Concerning the second assumption, it may be pointed out that many of the tendencies exhibited by current data must be regarded as statistically reversible. Falling birth rates due to drift of population to cities, to postponement of marriage on the part of professional classes, to the increasing cost of child culture, to the ubanization of rural life and to the restriction of immigration may be definitely altered by reversals in tendency. The flow of population may move into extraurban and subrural districts, where birth rates are more favorable to increase. The cost of child culture may, in part, be socially assumed. Improvement in economic conditions may lessen the drain on the resources of the family. The tendency for rural birth rates to fall may be checked. Immigration may increase with improving economic conditions. Death rates may be further reduced in many age classes and for many causes.

In fine, when we attempt to project into the future the components that

(24)

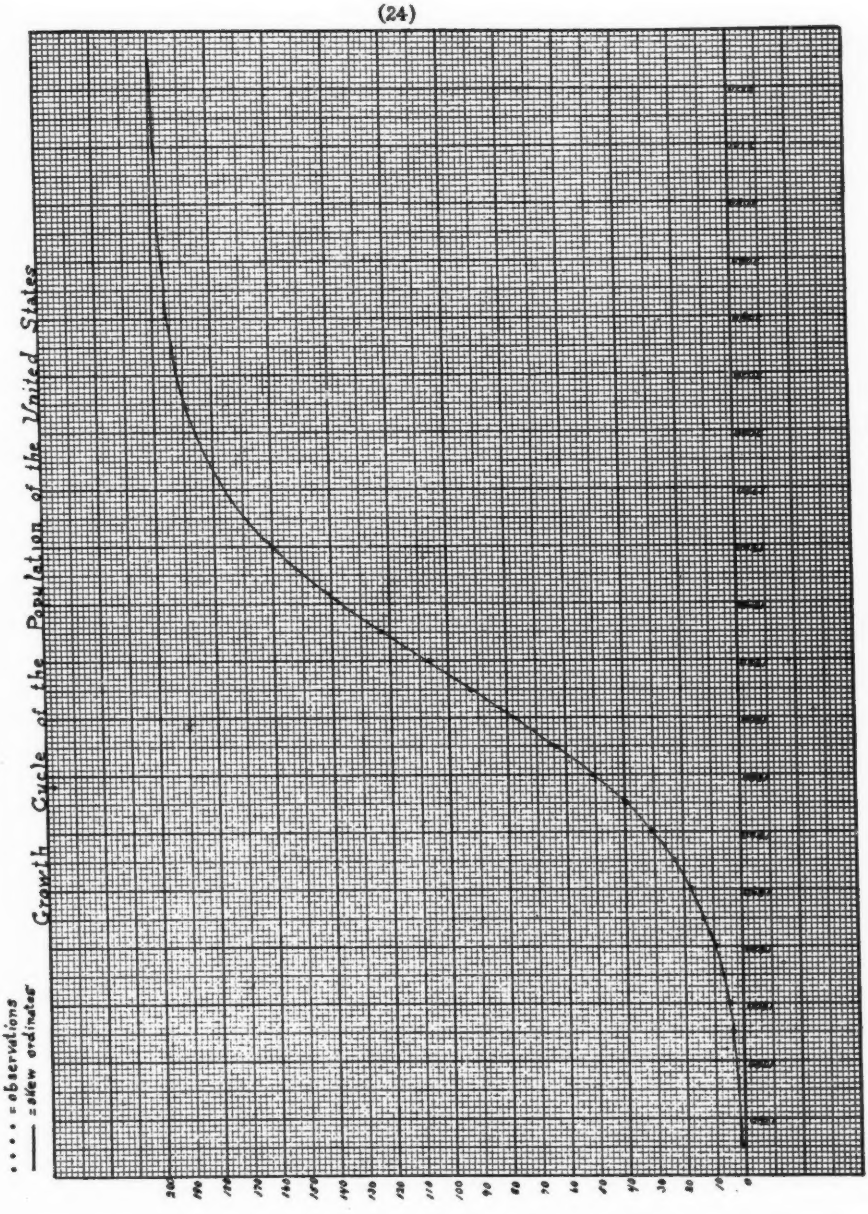
Growth Cycle of the Population of the United States
• • • • = observations
—— = skew ordinates

FIG. 1

determine the trend of natural increase, we encounter risks which vastly exceed those involved in the projection of the population series itself. Most of the data from which component trends must be determined cover but a brief period of time; while population data extends back for a century and a half. In this connection, it is not impertinent to inquire the criterion of relevance that will warrant a rejection of the items of the very series we are seeking to forecast.

It is a cardinal principle of logistic theory that the growth of population depends primarily on the continued supply of basic resources, physical and social, and that the dissipation of these resources is registered in the growth rate of the population itself. Any tendency of a population series toward skewness, that is, toward departure from the symmetric type of growth, is more likely to persist if it is systematic in character. The skew forms of the logistic function which we have developed permit us to measure any existing systematic tendency of the data toward skewness, and, therefore, to improve on the symmetric expectation of future growth.

In the case of the United States population, the evidence of skewness, insofar as it bears on the problem of expectation, is adverse to the conclusion that the ultimate limit of growth will be less than the symmetric asymptote. Conceding the light that the analysis of current tendencies may throw on the probable occurrence of future deviations from trend, the best criterion of long-time growth remains the logistic projection.

This statement, to be sure, does not relieve us of the necessity for recognizing the nature of the hazard that inheres in making a prediction from a trend extrapolation. The hazard involved in this type of inference arises from the assumption that the basic conditions of growth are stable, or, in other words, that the values of the parameters of the forecasting formula will remain substantially unchanged with the inclusion of new observations. Time alone can provide the final test of the continued validity of this assumption.

XV. **The Law of Organic Growth.** The law of organic growth in its most general form may be written:

$$p = L + H:[1 + e^{a+bt+s_1u_1+s_2u_2+s_3u_3}], \tag{44}$$

where $u_1 = \sin[m(t + q)]$; $u_2 = \log[1 + m^2(t + q)^2]$; $u_3 = \sqrt{1 + m^2(t + q)^2}$.

For most practical purposes, the evaluation of thirteen parameters is out of the question; hence, the restricted forms α, β, and γ, equation (18), will be the ones most generally employed.

I have made use of the term *law of organic growth* with reference to the logistic forms developed because I believe these functions to be the best means yet devised for the representation of the sequential changes which living organisms regularly manifest as individuals or societies. It states, in a quantitative form, all that is qualitatively implied by the so-called "law of diminishing returns" as this is commonly invoked by economists. The special sense in which I have used the term *law* may be expressed as follows:

A statistical law is a mathematical generalization on the behavior of a system of observations such that the implications of the formula are in accord with the assumptions basic to the phenomenon observed, and such that evaluations of the parameters of the formula determined from random samples are mutually consistent.

A statistical law, then, posits a system of relations manifesting itself in the form of observations which must be subjected to analysis before the true nature of their interrelations can be inferred. It expresses a probable, rather than a certain, inference; but, within the limitations of its claim to precision, it leaves reason no more free to reject its specification of reality than does a law of mechanics. Indeed, the point is still in dispute as to whether any law of science can be more than a statement of probabilities.

In contradistinction, the term *empirical formula* is properly restricted to cover the representation of the single set of observations at hand, and bears no necessary relation to any larger system. A sufficient test of an empirical formula is, therefore, the test of fit.

We may fit an indefinite number of formulas to a population series and obtain satisfactory results so far as agreement is concerned; but, on extrapolating, the same formulas will yield results that are patently absurd. The backward extrapolation for the population of the United States shown in Table V(b) represents the known facts as closely as could be expected when we take into consideration that census enumerations include aboriginal and immigrant populations as well as native born. Certainly, no random empirical formula, selected on the ground of goodness of fit, could be expected to yield as satisfactory a result.

Logistic theory does not, then, profess to guarantee infallibility of prediction. A population is not a mere aggregate of unrelated individuals inhabiting a restricted area, but a unified organization which grows by the utilization of total resources. When the supply of resources is profoundly disturbed or the basis of organizational unity destroyed, then the basis of prediction also is destroyed. And such reasoning is by no means peculiar to the sphere of social organization; for the integrity of any purely mechanical system is likewise conditioned by the assumption that the basis of coherence persists.

At this point, those in whom the speculative disposition is strong may query: if statistical prediction does not yield a certain result, is it, in the final analysis, superior to the ready and far less expensive method of guessing?

In answer, I can only say that, *a posteriori*, we can always, among a sufficiently large batch of guessers, find someone who has guessed well; but how, *a priori*, are we to know the good guesser from the poor? A population series consists of definite magnitudes, and any prediction of its development must result in the selection, out of a vast array of possible magnitudes, that which is most consistent with all the known facts. The gambler may elect to hazard his stake on the result of a random estimate; but the prudent will give heed to the exacting, if laborious, procedure of mathematical analysis.

ADDENDUM

Another solution of the theoretical problem stated in Section I may here be noted.

Given, as before, the function $y = f(x, a, b \cdots)$, we may, by assigning three approximate values to each parameter, compute 3^p sets of values for the function y, thus:

$$y_{11} = f(x, a_1b_1 \cdots); \qquad y_{12} = f(x, a_1b_2 \cdots); \qquad y_{13} = (x, a_1b_3 \cdots); \text{etc.}$$

From the observations Y, we may compute 3^p sets of the residuals $y - Y$; and from these several sets of residuals, the corresponding standard errors of estimate, σ, may be computed for each set of values of the function y; thus, we have:

$$\sigma_{11} = \phi(Y, x, a_1b_1)$$

$$\sigma_{12} = \phi(Y, x, a_1b_2)$$

$$\sigma_{13} = \phi(Y, x, a_1b_3)$$

Restricting the parameters to a, b, and holding a constant, we observe that the values $\sigma^2_{1\,1}$, $\sigma^2_{1\,2}$, $\sigma^2_{1\,3}$ must vary with the assigned values of the parameter b, and take a minimum value when b takes its true or most probable value. As the errors in the approximation to b increase positively and negatively without limit, the computed values of σ^2 will tend toward the infinite. They may, therefore, be assumed to lie on the arc of a parabola whose equation is a quadratic function of xa_1b; hence, we may form the following equations of representation:

$$\sigma^2_{1\,1} = k_{11} + l_{11}a_1 + m_{11}a_1^2.$$

$$\sigma^2_{1\,2} = k_{12} + l_{12}a_1 + m_{12}a_1^2.$$

$$\sigma^2_{1\,3} = k_{13} + l_{13}a_1 + m_{13}a_1^2.$$

By addition, we have,

$$\sigma^2_{1\,1} + \sigma^2_{1\,2} + \sigma^2_{1\,3} = k_{11} + k_{12} + k_{13} + (l_{11} + l_{12} + l_{13})a_1 + (m_{11} + m_{12} + m_{13})a_1^2.$$

By appropriate variations in subscript, similar equations may be written in a_2 and a_3, thus:

$$\sigma^2_{2\,1} + \sigma^2_{2\,2} + \sigma^2_{2\,3} = k_{21} + k_{22} + k_{23} + (l_{21} + l_{22} + l_{23})a_2 + (m_{21} + m_{22} + m_{23})a_2^2.$$

$$\sigma^2_{3\,1} + \sigma^2_{3\,2} + \sigma^2_{3\,3} = k_{31} + k_{32} + k_{33} + (l_{31} + l_{32} + l_{33})a_3 + (m_{31} + m_{32} + m_{33})a_3^2.$$

These three equations are all of the quadratic form, and may be conveniently written as follows:

$$A_1 = K_1 + L_1a_1 + M_1a_1^2.$$

$$A_2 = K_1 + L_1a_2 + M_1a_2^2.$$

$$A_3 = K_1 + L_1a_3 + M_1a_3^2.$$

By precisely similar reasoning, the following equations in b may be developed:

$$B_1 = K_2 + L_2 b_1 + M_2 b_1^2.$$

$$B_2 = K_2 + L_2 b_2 + M_2 b_2^2.$$

$$B_3 = K_2 + L_2 b_3 + M_2 b_3^2,$$

where

$$B_1 = \sigma_{11}^2 + \sigma_{21}^2 + \sigma_{31}^2; \quad B_2 = \sigma_{12}^2 + \sigma_{22}^2 + \sigma_{32}^2; \quad B_3 = \sigma_{13}^2 + \sigma_{23}^2 + \sigma_{33}^2.$$

Since the values of a_1, a_2, a_3 and b_1, b_2, b_3 are assigned, the two sets of equations may each be simultaneously solved to obtain values for K_1, L_1, M_1 and K_2, L_2, M_2. To obtain the conditions for A = a minimum, B = a minimum, we differentiate with respect to a and b, as follows:

$$D_a(A) = L_1 + 2M_1 a; \qquad D_b(B) = L_2 + 2M_2 b.$$

Setting these two equations equal to zero and solving, we obtain the adjusted values of a and b, thus:

$$a = -L_1 : 2M_1; \qquad b = -L_2 : 2M_2.$$

The extension of this method to the case of p parameters is obvious. Assigning three approximations to each parameter, we hold constant a value of one parameter (say a_1), we form all possible combinations of subscripts for the remaining parameters ($b_1 b_2 b_3$ with $c_1 c_2 c_3$ with etc.). This will yield 3^{p-1} values of σ^2, each of which is associated with a_1. Repeating this process, we can form similar sets of values of σ^2 by association with a_2 and a_3. We can then form the sums $A_1 = \sigma(Yxa_1bc \cdots)$; $A_2 = \sigma(Yxa_2bc \cdots)$; $A_3 = \sigma(Yxa_3bc \cdots)$. In all, $3 \times 3^{p-1}$ or 3^p distinct determinations of σ^2 will be required. In like manner, the equations for B_1, B_2, B_3 and C_1, C_2, C_3, etc. are formed. The solutions for the adjusted values $a, b, c, \cdots$ follow directly.

Since the method of solution given in Part I requires the computation of but 2^p values of σ^2, it is evident that the method of this section is the more onerous when considering the determination of a single set of adjusted values of parameters, the excess being of the order $3^p : 2^p = (1.5)^p$. However, being more precise, the present method will require fewer approximations to arrive at satisfactory values of the parameters sought. In other words, the mathematical advantage of economy lies with the theta technique; while the advantage of precision lies with the quadratic technique.

SELECTED BIBLIOGRAPHY

Dublin, L. I. and Lotka, A. J. On the True Rate of Natural Increase. J. A. S. A. S. 1925.

The True Rate of Natural Increase of the Population of the United States. Metron, Je. 1930.

Hotelling, H. Differential Equations Subject to Error and Population Estimates. J. A. S. A. 1927.

HOTELLING, H. AND F. Causes of Birth Rate Fluctuations. J. A. S. A. 1931.
KNIBBS, G. H. Laws of Growth of a Population. J. A. S. A. 1926–27.
LEHFELDT, R. A. The Normal Law of Progress; J. R. S. S. 1916.
LOTKA, A. J. Studies on the Mode of Growth of Material Aggregates. A. J. Sci. 1907.
PEARL, R. AND REED, L. J. On the Rate of Growth of the Population of the United States since 1790 and Its Mathematical Representation. Proc. N. A. Sci. 1920.
On the Mathematical Theory of Population Growth. Metron, 1923.
WILL, H. S. On Fitting Curves to Observational Series by the Method of Differences; Ann. M. S. My. 1930.
WOLFE, A. B. Is there a Biological Law of Human Population Growth? Q. J. Ec. Ag. 1927.
YULE, G. U. The Growth of Population and the Factors which Control It. J. R. S. S. Jn. 1924.

ON A METHOD FOR EVALUATING THE MOMENTS OF A BERNOULLI DISTRIBUTION[1]

By Everett H. Larguier, S.J.

1. The moments (per unit frequency) of a frequency distribution have long been regarded as useful characteristics of the distribution. If we denote the moment about the arithmetic mean by μ, we have for the Bernoulli distribution

$$\mu_s = \sum_{x=0}^{n} (\bar{x})^s f(x),$$

where $\bar{x} = x - np$ and $f(x) = \binom{n}{x} p^x q^{n-x}$.

To evaluate the s-th moment about the arithmetic mean has always been a laborious task. Karl Pearson[2] gave the s-th moment about the arithmetic mean as,

$$\mu_s = \left[\frac{d^s}{dx^s}[qe^{px} + pe^{-qx}]^n\right]_{x=0}, \tag{1}$$

which he said at that time was perhaps the easiest expression for obtaining these moment coefficients by successive differentiation. Romanovsky,[3] however, was able to develop the recursion formula,

$$\mu_{s+1} = pq\left[ns\mu_{s-1} + \frac{d\mu_s}{dp}\right], \tag{2}$$

for the moments about the mean. Another relation for these moments is

$$\mu_{s+1} = \sum_{i=0}^{s-1} \binom{s}{i} [npq\mu_i - p\mu_{i+1}]. \tag{3}$$

Recently Kirkham[4] gave the expressions for the first eight moments which, however, are not in a form well adapted for numerical calculation on a machine.

---

[1] Presented to the American Mathematical Society, January 2, 1936.

[2] Karl Pearson, *Biometrika*, vol. 12 (1918–1919), footnote, p. 270. This expression is obtained from the moment-generating function. Obviously this method is exceedingly impractical for numerical calculations.

[3] V. Romanovsky, "Note on the moments of the binomial $(p + q)^n$ about its mean," *Biometrika*, vol. 15 (1923). Recently this expression was given a simple proof by A. T. Craig (*Bull. Amer. Math. Soc.*, vol. 40, pp. 262–264) and extended to the Poisson case.

[4] W. J. Kirkham, "Moments about the arithmetic mean of a binomial frequency distribution," *Annals of Mathematical Statistics*, vol. VI, pp. 96–101.

2. It is the purpose of this paper to express the s-th moment about the arithmetic mean in the form

$$\mu_s = \sum_{t=1}^{t=s} F_{s,t}(n)p^t, \tag{4}$$

where $F_{s,t}(n)$ are determinable functions of n dependent on s and t. We note here that p and q are the probabilities of the success and failure of an event in a single trial.

Since we know that $\mu_2 = npq$ and $\mu_1 = 0$, it is evident that the part of (2) enclosed in [ ] will be of degree 2 less than $s + 1$ in p and hence (4) will satisfy as a representation of the moment.

3. To obtain a recursion formula for the functions $F_{s,t}(n)$ we differentiate (4) with respect to p. This gives

$$\frac{d\mu_s}{dp} = \sum_{t=1}^{s} tF_{s,t}(n)p^{t-1}.$$

By (2) we may then write

$$\begin{aligned}\sum_{t=1}^{s+1} F_{s+1,t}(n)p^t &= p(1-p)ns\sum_{t=1}^{s-1} F_{s-1,t}(n)p^t + p(1-p)\sum_{t=1}^{s} tF_{s,t}(n)p^{t-1}\\ &= ns\sum_{t=2}^{s} F_{s-1,t-1}(n)p^t - ns\sum_{t=3}^{s+1} F_{s-1,t-2}(n)p^t\\ &\qquad + \sum_{t=1}^{s} tF_{s,t}(n)p^t - \sum_{t=2}^{s+1}(t-1)F_{s,t-1}(n)p^t.\end{aligned}$$

Since this is an identity in p, we have immediately the following recursion formula for determining $F_{s,t}(n)$:

$$\begin{aligned}F_{s,t}(n) = n(s-1)F_{s-2,t-1}(n) - n(s-1)F_{s-2,t-2}(n) + tF_{s-1,t}(n)\\ - (t-1)F_{s-1,t-1}(n)\end{aligned} \tag{5}$$

in which

$$F_{0,0}(n) = 1; \quad \text{and} \quad F_{s,t}(n) = 0 \quad \text{for} \quad \begin{cases} t > s; \\ t < 1,\ s > 0; \\ t = 1,\ s = 1. \end{cases} \tag{6}$$

These definitions arise from the known values of the moments and the conditions imposed by the identity in p.

By means of (5) and (6) we are able to obtain very readily the values for $F_{s,t}(n)$ which are given in Table 1.

TABLE I

Values of $F_{s,t}(n)$

s	$F_{s,1}(n)$	$F_{s,2}(n)$	$F_{s,3}(n)$	$F_{s,4}(n)$
1	0	0	0	0
2	n	$-n$	0	0
3	n	$-3n$	$2n$	0
4	n	$-7n + 3n^2$	$12n - 6n^2$	$-6n + 3n^2$
5	n	$-15n + 10n^2$	$50n - 40n^2$	$-60n + 50n^2$
6	n	$-31n + 25n^2$	$180n - 180n^2 + 15n^3$	$-390n + 415n^2 - 45n^3$
7	n	$-63n + 56n^2$	$602n - 686n^2 + 105n^3$	$-2100n + 2590n^2 - 525n^3$
8	n	$-127n + 119n^2$	$1932n - 2394n^2 + 490n^3$	$-10206n + 13895n^2 - 3850n^3 + 105n^4$

s	$F_{s,5}(n)$	$F_{s,6}(n)$
1	0	0
2	0	0
3	0	0
4	0	0
5	$24n - 20n^2$	0
6	$360n - 390n^2 + 45n^3$	$-120n + 130n^2 - 15n^3$
7	$3360n - 4270n^2 + 945n^3$	$-2520n + 3234n^2 - 735n^3$
8	$25200n - 35700n^2 + 10990n^3 - 420n^4$	$-31920n + 46004n^2 - 14770n^3 + 630n^4$

s	$F_{s,7}(n)$	$F_{s,8}(n)$
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	$720n - 924n^2 + 210n^3$	0
8	$20160n - 29232n^2 + 9520n^3 - 420n^4$	$-5040n + 7308n^2 - 2380n^3 + 105n^4$

With this table it is a relatively easy task to evaluate the first eight moments with the aid of a calculating machine.

4. As an illustration of the preceding we propose to evaluate the first eight moments about the arithmetic mean for the binomial, $(.06785 + .93215)^{378}$. We first evaluate the coefficients $F_{s,t}(n)$.

TABLE II[5]

Values of $F_{s,t}(378)$

s	$F_{s,1}(378)$	$F_{s,2}(378)$	$F_{s,3}(378)$	$F_{s,4}(378)$	$F_{s,5}(378)$
1	0	0	0	0	0
2	378	−378	0	0	0
3	378	−1,134	756	0	0
4	378	426,006	−852,768	426,384	0
5	378	1,423,170	−5,696,460	7,121,520	−2,848,608
6	378	3,560,382	784,501,200	−2,371,307,400	2,374,868,160
7	378	7,977,690	5,573,275,090	−27,986,054,000	50,430,749,000
8	378	16,955,190	26,123,640,500	1,937,705,370,000	−7,986,171,610,000

s	$F_{s,6}(378)$	$F_{s,7}(378)$	$F_{s,8}(378)$
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	−791,622,720	0	0
7	−39,236,327,400	11,210,379,300	0
8	12,070,808,800,000	−8,064,644,270,000	2,016,161,070,000

Then running off the powers of p, we have:

$$p = .067\ 85 \qquad p^5 = .000\ 001\ 437\ 968\ 13$$
$$p^2 = .004\ 603\ 622\ 5 \qquad p^6 = .000\ 000\ 097\ 566\ 137\ 6$$
$$p^3 = .000\ 312\ 355\ 787 \qquad p^7 = .000\ 000\ 006\ 619\ 862\ 44$$
$$p^4 = .000\ 021\ 193\ 340\ 1 \qquad p^8 = .000\ 000\ 000\ 449\ 157\ 667$$

Applying (4) we have

[5] In this table, as well as in the one that follows, all values are correct to nine significant figures.

TABLE III

Values of $p^t F_{s,t}(378)$

s	2	3	4	5
$pF_{s,1}(378)$	25.6473	25.6473	25.6473	25.6473
$p^2F_{s,2}(378)$	−1.7401693	−5.2205079	1961.17087	6551.73743
$p^3F_{s,3}(378)$	0.	.2361410	−266.36702	−1779.32225
$p^4F_{s,4}(378)$	0.	0.	9.03650	150.92880
$p^5F_{s,5}(378)$	0.	0.	0.	−4.09621
$p^6F_{s,6}(378)$	0.	0.	0.	0.
$p^7F_{s,7}(378)$	0.	0.	0.	0.
$p^8F_{s,8}(378)$	0.	0.	0.	0.
μ_s	23.9071307	20.6629331	1729.48765	4944.89507

s	6	7	8
$pF_{s,1}(378)$	25.647	25.65	25.6
$p^2F_{s,2}(378)$	16390.655	36726.27	78055.3
$p^3F_{s,3}(378)$	245043.490	1740844.73	8159870.3
$p^4F_{s,4}(378)$	−50255.924	−593117.96	41066448.9
$p^5F_{s,5}(378)$	3414.985	72517.81	−11483860.3
$p^6F_{s,6}(378)$	−77.236	−3828.14	1177702.2
$p^7F_{s,7}(378)$	0.	74.21	−53386.8
$p^8F_{s,8}(378)$	0.	0.	905.6
μ_s	214541.617	1253242.57	38945760.8

This gives us the desired moments about the arithmetic mean of the binomial $(.06785 + .93215)^{378}$. These values may be rapidly checked by applying (3) to μ_8.

SAINT LOUIS UNIVERSITY,
SAINT LOUIS, MISSOURI.

A METHOD OF DETERMINING THE REGRESSION CURVE WHEN THE MARGINAL DISTRIBUTION IS OF THE NORMAL LOGARITHMIC TYPE

By Carl-Erik Quensel

Assistant at the Statistical Institute of the University of Lund, Sweden

In a paper[1] in this Journal Professor S. D. Wicksell gave the general outlines of a new method of calculating the regression lines. This problem was later on treated in detail by Dr. Walter Andersson.[2] His method was to develop the formulas for the regression lines into a series of orthogonal polynomials under the assumption that the marginal distribution of the independent variate belonged to certain mathematically defined distributions, and to determine the constants with the aid of the method of the least squares.

Among other cases he treated also the case where the marginal distribution was of the normal logarithmic type:

$$F(x) = \frac{\log e}{\sigma_l \sqrt{2\pi}\,(x - a)} e^{-\frac{1}{2}\left[\frac{\log(x-a) - l}{\sigma_l}\right]^2}. \tag{1}$$

But as his method is entirely different from the method I shall give here, I will not go any further into the method used by Dr. Andersson.

When the correlation surface $F(x, y)$ of the variates x and y is given and then of course also the marginal distribution of x, $F(x)$, it is known that the mean y_x of the dependent variate y in an infinitely small array with the value of x between x and $x + dx$ is given as a function of the independent variate x by the following formula (2)

$$y_x = \frac{\int yF(x, y)\, dy}{\int F(x, y)\, dy}. \tag{2}$$

In this formula the integrals are to be extended over the whole domain of the variation of y.

If now we make any transformation of x by introducing a new variate u, related to x by the formula $u = \psi(x)$, where we must suppose that u is a one-valued function of x and contrary, the distribution $f(u, y)$ of the variate u and y is given by the relation

$$f(u, y)\, du\, dy = F(x, y)\, dx\, dy \tag{3}$$

[1] S. D. Wicksell. Remarks on Regression. Annals of Mathematical Statistics, 1930.

[2] Walter Andersson. Researches into the theory of Regression. Meddelande från Lunds Astronomiska Observatorium. Ser II. N:r 64.

Writing the formula (2) in the following form:

$$y_x = \frac{\int yF(x, y)\, dx\, dy}{\int F(x, y)\, dx\, dy};$$

we see at once that the mean y_x can be given as the following function of u:

$$y_x = \frac{\int yf(u, y)\, dy}{\int f(u, y)\, dy}. \tag{4}$$

This relation, of course, is self-evident. The mean of the dependent variate in an array of the independent variate will be unchanged, when we change the variate x for another variate u, related to x by a one-valued function.

The problem of finding the regression line of the mean y_x can in such a way be much simplified, if it is possible to make a favorable transformation of the independent variate x.

As shown by Professor Wicksell[3] we may, under certain conditions concerning the marginal distribution $f(u)$, write the expression of the regression line in the following form:

$$y_x = \sum_0^\infty (-1)^n \frac{\lambda_{n,1}}{n!} \frac{f^{(n)}(u)}{f(u)}; \tag{5}$$

where the $\lambda_{n,1}$ coefficients are the seminvariants of the distribution of u and y.

The conditions which the function $f(u)$ must satisfy are among others that the function and all its derivates are continuous in the domain of variation and that the function and its derivates disappear in the limits of that domain. These conditions are satisfied by the normal curve of error.

In the case where the distribution of u is normal, the derivates $f^{(n)}(u)$ take the following form:

$$f^{(n)}(u) = (-1)^n H_n(u) f(u); \tag{6}$$

where the expressions $H_n(u)$ are the well known Hermitian polynomials.

The formula (5) takes the following simple form.

$$y_x = \sum_0^\infty \frac{\lambda_{n,1}}{n!} H_n(u) \tag{7}$$

If we can change the given marginal distribution $F(x)$ by a favorable substitution $u = \psi(x)$ into a normal curve, and if, this substitution made, we can

---

[3] S. D. Wicksell. Analytical Theory of Regression. Meddelande från Lunds Astronomiska Observatorium. Ser II. N:r 69.

calculate the coefficients $\lambda_{n,1}$ from the moments or other known characteristics of the given correlation distribution, $F(x, y)$, it is possible to express the regression line as the formula (8) shows:

$$y_x = \sum_0^\infty \frac{\lambda_{n,1}}{n!} H_n[\psi(x)] \tag{8}$$

It must be observed that the polynomials $H_n[\psi(x)]$ are orthogonal with regard to the distribution $F(x)$ of the independent variate x. We have

$$\int H_i[\psi(x)]H_j[\psi(x)]F(x)\,dx = \int H_i(u)H_j(u)f(u)\,du = 0 \qquad i \neq j$$

Not in all cases it will perhaps be possible to calculate the $\lambda_{n,1}$ coefficients, when we have transformed the marginal distribution into the normal curve, but in one case it is rather simple to calculate these coefficients from the moments given.

The case alluded to is the one, where the variate u is given from x by the relation $u = \log(x - a)$, that is that the marginal distribution is of the so called normal logarithmic type (1).

In that case it is possible to calculate the $\lambda_{n,1}$ coefficients from the marginal moments $V_{n,0}$ and from the correlation moments of the type $V_{n,1}$.

We suppose that the marginal distribution is of the logarithmic type and that from the moments of the x distribution we have determined the three constants a, σ_l and l in the usual manner.[4]

Then we calculate from the given correlation distribution the moments $V'_{n,0}$ about the point $x = a$ and the correlation moments $V'_{n,1}$ about the point $x = a$ and $y = m_y$ (the mean value of the y-variate).

From these moments it is possible to calculate the $\lambda_{n,1}$ coefficients in the following way.

The characteristic function of u and y is given by the following relation:

$$U(t_1t_2) = e^{\sum \frac{\lambda_{kl}}{k!\,l!} t_1^k t_2^l} = \iint e^{t_1u+t_2y} f(u, y)\,du\,dy \tag{9}$$

where the integrals are extended over the whole domain of variation.

If the distribution of u is according to the normal law, we have $\lambda_{k,0} = 0$ for $k \geqq 3$, but in the calculations here it is not at all necessary to suppose anything about these higher seminvariants. On the other side, the correlation distribution $f(u, y)$ is obtained from the characteristic function by the inversion theorem.

$$f(u, y) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\sum \frac{\lambda_{kl}}{k!\,l!}(i\,w_1)^k(i\,w_2)^l} e^{-iw_1u-iw_2y}\,dw_1\,dw_2 \tag{10}$$

[4] How these are to be determined is shown in Pae- Tsi Yuan. On the logarithmic Frequency distribution and the Semi-Logarithmic Correlation Surface. Annals of Mathematical Statistics, 1933.

But we can also get the following relation

$$\int e^{t_1 u} f(u, y)\, du = \frac{1}{2\pi} \int e^{-iw_2 y} e^{\sum \frac{\lambda_{kl}}{k!\,l!} t_1^k (i w_2)^l} dw_2 \tag{11}$$

Of this last expression (11) between the characteristic function and the distribution function I will make use in the following.

The moments V'_{ij} of the distribution $F(x, y)$ about the point $x = a$, $y = m_y$ are given by the formula

$$V'_{ij} = \int\int (x - a)^i (y - m_y)^j F(x, y)\, dx\, dy. \tag{12}$$

If we write y instead of $y - m_y$ and instead of $x - a$ we write e^{bu} ($b = {}_{10}\log e$) the expression (12) takes the following form:

$$V'_{ij} = \int\int e^{ibu} y^j f(u, y)\, du\, dy \tag{13}$$

For the marginal moments of x about the point $x = a$ we get

$$V'_{n,0} = \int_a^\infty (x - a)^n F(x)\, dx = \int_{-\infty}^\infty e^{nbu} f(u)\, du \tag{14}$$

Comparing this formula (14) with the expression for the characteristic function of the distribution $f(u)$

$$U(t_1) = \int_{-\infty}^\infty e^{t_1 u} f(u)\, du = e^{\sum \frac{\lambda_{k,0}}{k!} t_1^k}; \tag{15}$$

we find the following simple relation

$$V'_{n,0} = e^{\sum \frac{\lambda_{k,0}}{k!} (n b)^k} \tag{16}$$

For the moments of the type $V'_{n,1}$ we get

$$V'_{n,1} = \int\int e^{nbu} y f(u, y)\, du\, dy = \int y dy \int e^{nbu} f(u, y)\, du. \tag{17}$$

If we compare the last integral in the formula (17) $\int e^{nbu} f(u, y) du$ with the formula (11) we see that we can write (17) as follows:

$$V'_{n,1} = \frac{1}{2\pi} \int y dy \int e^{-iw_2 y} e^{\sum \frac{\lambda_{kl}}{k!\,l!} (n b)^k (i w_2)} dw_2 \tag{18}$$

From the sum $\sum \frac{\lambda_{kl}}{k!l!} (nb)^k (iw_2)^l$ we may take out the part $\sum \frac{\lambda_{k,0}}{k!} (nb)^k$,

where l is zero and which therefore does not contain any dignity of w_2, and write the remainder in the following form:

$$\sum \frac{\lambda'_l}{l!}(iw_2)^l$$

where we have

$$\lambda'_1 = \lambda_{11}nb + \frac{\lambda_{21}}{2!}(nb)^2 + \frac{\lambda_{31}}{3!}(nb)^3 \cdots$$

$$\frac{\lambda'_2}{2!} = \frac{\lambda_{02}}{2!} + \frac{3\lambda_{12}}{3!}nb + \frac{6\lambda_{22}}{4!}(nb)^2 \cdots$$

The integral $\frac{1}{2\pi}\int e^{-iw_2y}e^{\sum \frac{\lambda'_l}{l!}(iw_2)^l}$ may be considered as a frequency distribution $\varphi(y)$ with the seminvariants λ'_l.

The formula (18) will thus be written

$$V'_{n,1} = e^{\sum \frac{\lambda_{k0}}{k!}(nb)^k}\int ydy\,\varphi(y) \tag{19}$$

According to (16) we have

$$e^{\sum \frac{\lambda_{k0}}{k!}(nb)^k} = V'_{n,0}$$

and as

$$\int ydy\,\varphi(y) = \lambda'_1 = \lambda_{11}nb + \frac{\lambda_{21}}{2!}(nb)^2 + \frac{\lambda_{31}}{3!}(nb)^3 \cdots$$

we get

$$V'_{n,1} = V'_{n,0}\cdot\lambda'_1 \tag{20}$$

or

$$\frac{V'_{n,1}}{V'_{n,0}} = \lambda_{11}nb + \frac{\lambda_{21}}{2!}(nb)^2 + \frac{\lambda_{31}}{3!}(nb)^3 \cdots \tag{21}$$

We see that in the formulas for $V'_{n,1}$ we have all the seminvariants $\lambda_{n,1}$ involved. A successive determination of the seminvariants $\lambda_{n,1}$ with the aid of the moments of the same and lower degree is therefore not possible.

However, when we use the formula (8) for the regression, we must suppose that the seminvariants $\lambda_{n,1}$ with growing n converge rather soon towards zero.

If the successive differences $\Delta^n\left(\frac{V'_{n,1}}{V'_{n,0}}\right)$ of the quotients $\frac{V'_{n,1}}{V'_{n,0}}$ are calculated, it may be possible to judge, how far it is possible to go with success. These differences will in most cases diminish rather soon and we shall therefore in most cases get a value of n about which we can suppose that the differences of higher order than this will all be so small that they can be neglected and as a consequence of this fact all higher seminvariants can be neglected too.

When this value of n has been determined, the n first seminvariants will all be obtained from the n first quotients $\frac{V'_{n,1}}{V'_{n,0}}$.

Thus we finally get the regression line as follows:

$$y_x = m_y + \sum_{1}^{n} \frac{\lambda_{i,1}}{i!} H_i[\log (x - a) - l]$$

or in standardized units:

$$y_x = m_y + \sum_{1}^{n} \frac{\lambda_{i,1}}{i!\,\sigma_l^i} H_i\left[\frac{\log (x - a) - l}{\sigma_l}\right]$$

THE STANDARD ERROR OF A "SOCIAL FORCE"

By Stuart C. Dodd

I. Definitions

In the theory of measurement of social forces certain special cases of frequent occurrence where the population shifts from one date of measurement to the next require the derivation of appropriate standard error formulae.

The theory may be briefly restated[1] in equations as follows: any measurable social change, C, in a population, P, may be defined as the difference in mean scores, S, from surveys or measurements on the dates denoted by subscripts

$$C_{2-1} = S_2 - S_1 = \frac{\Sigma s_2}{P} - \frac{\Sigma s_1}{P} \tag{1}$$

The momentum of a social change may be defined as the product of its time rate in years and the population that is being changed

$$M_{2-1} = PV_{2-1} \tag{2}$$

$$= \frac{PC_{2-1}}{Y_{2-1}} = \frac{P}{Y_{2-1}}(S_2 - S_1) \tag{2a}$$

where Y_{2-1} is the period from date 1 to date 2 and V is the velocity, or speed of change, in that period. The acceleration of a social change is definable as the rate of change of the velocity of change

$$A = \frac{V_{4-3} - V_{2-1}}{.5Y_{(4-3-2+1)}} \tag{3}$$

where each velocity, being an average for its period, is taken as representing the mid-date of that period.

The resultant social force which produces a measured change is now definable as that which accelerates the change in a population. It is measurable as the product of the acceleration and the population.[2]

$$F = AP \tag{4}$$

$$= \frac{P}{.5Y_{(4-3-2+1)}}\left(\frac{S_1}{Y_{2-1}} - \frac{S_2}{Y_{2-1}} - \frac{S_3}{Y_{4-3}} + \frac{S_4}{Y_{4-3}}\right) \tag{5}$$

---

[1] *A Controlled Experiment on Rural Hygiene in Syria*, Dodd, S. C., Publications of the American University of Beirut, Syria, Social Science Series No. 7, 1934, pp. 336.

Also, *A Theory for the Measurement of Some Social Forces*, Dodd, S. C., Scientific Monthly, Vol. XLIII, No. 1, July 1936, pp. 58-62

[2] Force thus defined in terms of its effect is a resultant force, i.e., the residual force after deducting all resisting forces from the total force in the direction of the change observed. This formula defines quantitatively and exactly the "net" force not the "gross" force

II. The Sampling error of one case (momentum)

The formulae for the standard errors of sampling for the above concepts, social change, velocity, momentum, acceleration and force, (C, V, M, A, and F) have been published for the case where the population, P, is the same on all dates of measurement. But it is not always possible to observe the ideal experimental technic of holding the population unchanged in number nor to select out individuals common to all the surveys and to neglect the rest. Ordinarily there will be different P's, P_1, P_2, P_3, and P_4, at the different dates.

To derive the standard errors of (2) and (4) when P shifts, each P is considered to be a sub-sample[3] of the main sample which is $(P_1 + P_2 + P_3 + P_4)$. The orthodox view of sampling is taken where the sub-samples may differ in size but maintain fixed proportions in each main sample which is drawn from the "parent" population.

Let primes denote an M, or other function of (1) to (5), which is an approximation due to the shifting of the population and the use of an average P.

To simplify and generalize the notation, let k denote the constant term compounded of P's and Y's which is associated with each S. The first subscript of k denotes the function, f, which is any particular one of the left hand members of equations (1) to (5) and the second subscript denotes the date of its S. Thus, from (2a)

$$k_{M1} = \frac{-P_1 + P_2}{2Y_{2-1}} = -k_{M2} \tag{6}$$

Then (2) may be rewritten:

$$M'_{2-1} = S_1 k_{M1} + S_2 k_{M2} \tag{7}$$

$$= \sum_1^2 Sk_M. \tag{7a}$$

To derive the standard error of (7) the total differential is:

$$dM'_{2-1} = k_{M1}\, d\left(\frac{\Sigma s_1}{P_1}\right) + k_{M2}\, d\left(\frac{\Sigma s_2}{P_2}\right) \tag{8}$$

If Q_{12} denotes the population common to both dates of measurement so that:

$$\begin{aligned} P_1 &= Q_{12} + Q_1 \\ P_2 &= Q_{12} + Q_2 \end{aligned} \tag{9}$$

producing the change. It thus measures only the *observable part* of the total forces in the situation. The fundamental problem remains, as always in science, to observe more adequately, to devise experimental and statistical technics for measuring the different forces (in isolation and in combinations) which facilitate or resist the measured change.

[3] The author is indebted to Mr. S. S. Wilks (Princeton) for this method of deriving these standard errors in a fluctuating population.

and, since the differential of a sum is the sum of the differentials of the several terms, (8) becomes

$$dM'_{2-1} = \frac{k_{M1}}{P_1}\left(\sum_1^{Q_{12}} ds_1 + \sum_1^{Q_1} ds_1\right) + \frac{k_{M2}}{P_2}\left(\sum_1^{Q_{12}} ds_2 + \sum_1^{Q_2} ds_2\right) \tag{10}$$

Squaring gives

$$\begin{aligned} (dM'_{2-1})^2 &= \frac{k^2_{M1}}{P^2_1}\left(\sum ds_1\right)^2 + \frac{k^2_{M2}}{P^2_2}\left(\sum ds_2\right)^2 \\ &+ \frac{2\,k_{M1}\,k_{M2}}{P_1\,P_2}\left[\sum_1^{Q_{12}} ds_1 \sum_1^{Q_{12}} ds_2 + \sum_1^{Q_{12}} ds_1 \sum_1^{Q_2} ds_2 + \sum_1^{Q_1} ds_1 \sum_1^{Q_{12}} ds_2 + \sum_1^{Q_1} ds_1 \sum_1^{Q_2} ds_2\right] \end{aligned} \tag{11}$$

On summing and dividing by the number of cases to get the expected values, the last three terms in the square brackets vanish. Using the relation where, in random sampling, the correlation between two variables is the same as the correlation between their means

$$r_{12} = r_{\bar{s}_1\bar{s}_2} = \frac{\Sigma\, S_1S_2}{Q_{12}\,\sigma_1\sigma_2} = \frac{\Sigma\left(\frac{\Sigma\, s_1}{Q_{12}}\cdot\frac{\Sigma\, s_2}{Q_{12}}\right)}{Q_{12}\,\frac{\sigma_1\,\sigma_2}{\sqrt{Q_{12}\cdot Q_{12}}}} \tag{12}$$

gives

$$\sigma^2_{M'_{2-1}} = \frac{k^2_{M1}\,\sigma^2_1}{P_1} + \frac{k^2_{M2}\,\sigma^2_2}{P_2} + \frac{2\,k_{M1}\,k_{M2}\,Q_{12}\,\sigma_1\,\sigma_2\,r_{12}}{P_1\,P_2} \tag{13}$$

Standard error of momentum when the population shifts

The best estimates of σ_1 and σ_2 are the standard deviations of the scores, s_1 and s_2, and the best estimate of r_{12} is, strictly, the covariance of the common cases divided by the two sigmas. Unless the selection of Q_{12} out of P_1 and P_2 curtails the range in some way (i.e., Q_{12} is not a random selection), then, except for sampling variation, σ_1 and σ_2 are the same in the Q_{12} population as in the P_1 and P_2 populations so that there is only a sampling discrepancy between the ratio above and the r_{12}, the observed correlation between the s_1 and s_2 scores in the Q_{12} population.

III. The generalized standard error

The above standard error may be readily generalized. Any of the equations (1) to (5) may be expressed as a simple linear sum of the products of a variable, S, and its appropriate constant, k.

$$f = \sum_{i=1}^{i=n} S_i\,k_{fi} \tag{14}$$

where f is any one of the concepts S, C, V, A, M or F defined by (1) to (5) and n is the number of surveys, or different S's involved, and i denotes each survey in turn from 1 to n. Thus where f means F, (5) becomes:

$$\begin{aligned} f'_F = F' &= k_{F1} S_1 + k_{F2} S_2 + k_{F3} S_3 + k_{F4} S_4 \\ &= \sum_{i=1}^{i=4} k_{Fi} S_i \end{aligned} \tag{15}$$

where

$$k_{F1} = -\,k_{F2} = \frac{P_1 + P_2 + P_3 + P_4}{2\, Y_{(4-3-2+1)}\, Y_{(2-1)}} \tag{16a}$$

$$k_{F4} = -\,k_{F3} = \frac{P_1 + P_2 + P_3 + P_4}{2\, Y_{(4-3-2+1)}\, Y_{(4-3)}}. \tag{16b}$$

In the special case when a force, F, has been determined from only three surveys using two consecutive periods, $n = 3$ and

$$k_{1F} = \frac{P_1 + P_2 + P_3}{1.5\, Y_{(3-1)}\ (Y_{2-1})} \tag{16c}$$

$$k_{F2} = -\,\frac{(P_1 + P_2 + P_3)\,(Y_{(2-1)} + Y_{(3-2)})}{1.5\, Y_{(3-1)}\, Y_{(3-2)}\, Y_{(2-1)}} \tag{16d}$$

$$k_{F3} = \frac{P_1 + P_2 + P_3}{1.5\, Y_{(3-1)}\, Y_{(3-2)}} \tag{16e}$$

If the difference between two forces (or other functions, f) has been measured in either the same or in different populations and the significance of the difference in terms of its standard error is desired, f of (14) can also denote that difference.

$$f_{dF} = F_a - F_b; \qquad f_{dM} = M_a - M_b; \text{etc.} \tag{17}$$

It is only necessary to write the difference as a linear sum of products of S and k on the model of (2a) or (5) to get the k-values for that particular f.

It is now possible to write the standard error formula for f in a single generalized form that covers all the concepts and their differences as defined in equations (1) to (5), (14) and (17). Observing that (14) is the general case for n surveys of the particular case (7a) where $n = 2$, it becomes evident, that on taking differentials, squaring, summing, and dividing the linear sum of the n terms of (14) there results n^2 terms of which there are n that are variances (times constants) of the sort $\frac{k^2\sigma^2}{P}$ and $\frac{n^2 - n}{2}$ are different terms each occurring twice that are covariances (times constants) of the sort $\frac{kkQ\sigma\sigma r}{PP}$. From these

rough considerations as well as from rigorous derivation, the generalized standard error of (14) is found to be:

$$\sigma_f^2 = \sum_1^{n^2} \frac{k_{fi}\,\sigma_i\,k_{fj}\,\sigma_j\,Q_{ij}\,r_{ij}}{P_i\,P_j}, \tag{18}$$

The generalized standard error.

Where i and j denote each of the n surveys in turn. There will thus be n^2 terms to be summed—the number of combinations of i with j including the cases where $i = j$.

The derivation of (18) as well as its computation from data and its interpretation in special cases can all be made clearer by arranging the terms in a square array as follows:

	$i \rightarrow$	1	2	......	n
j $\downarrow$	Coefficients $\downarrow$ $\rightarrow$	$\frac{k_{f1}\sigma_1}{P_1}$	$\frac{k_{f2}\sigma_2}{P_2}$	......	$\frac{k_{fn}\sigma_n}{P_n}$
1	$\frac{k_{f1}\sigma_1}{P_1}$	P_1 ( )	$Q_{12}r_{12}$ ( )	......	$Q_{1n}r_{1n}$ ( )
2	$\frac{k_{f2}\sigma_2}{P_2}$	$Q_{12}r_{12}$ ( )	P_2 ( )	......	$Q_{2n}r_{2n}$ ( )
$\vdots$	$\vdots$	$\vdots$	$\vdots$	......	$\vdots$
n	$\frac{k_{fn}\sigma_n}{P_n}$	$Q_{1n}r_{1n}$ ( )	$Q_{2n}r_{2n}$ ( )	......	P_n ( )

To get σ_f write the computed values of the coefficients $\frac{k\sigma}{P}$ as captions of rows and of columns and write each computed Qr value in its appropriate cell, noting that in the main diagonal cells the self-correlations are unities and the population common to both column and row surveys, Q_{ii} is the entire population of that survey as $Q_{ij} = P_i$ when $i = j$. Thus $Q_{11} = P_1$. Next in each cell's parenthesis enter the product of three factors, namely: a) the cell Qr term, b) the column coefficient, and c) the row coefficient. The sum of these products in the parentheses, n^2 in number, is σ_f^2 of (18).

From the above square array it becomes clear that whenever in (17) the difference of two observed forces, or other functions, is derived from *different* populations the Q between these populations is zero so that the entire product terms in those cells vanish. Thus in the very simplest and familiar case of

comparing two means from different populations, $n = 2$, $Q_{12} = 0$, $k = 1$, and (18) reduces to the usual sum of the two variances of the two means

$$\sigma^2 \text{ difference in means} = \frac{\sigma_1^2}{P_1} + \frac{\sigma_2^2}{P_2} \tag{19}$$

IV. Some special cases

It should be observed that the above formulae for the standard errors when P shifts all become identical with the simpler formulae previously derived for the case of a constant P. In this case, every $Q_{pq} = P_p = P_q$ and in the square array (in addition to k's which no longer involve an average P), the Q or P of the cells and the P's in the row coefficients, may be omitted as they cancel each other out.

Another special but very frequent case is where the social change is not given in terms of a difference in means, S_1 and S_2, but in terms of a difference in percentages, as when a literacy rate rises from 30% to 40%. A percentage can be viewed as a mean of a two-category, all-or-none, present-or-absent variable such as: A, non-A (foreign or native born, literate or illiterate, etc), where A is assigned a value of 1 and non-A a value of 0. Then the sum of the values of A, each times its frequency, divided by the population is both a proportion and a mean. Its standard error in the percentage, p, form of expression is then equal to it in the mean form:

$$\sigma_p = \frac{p\sqrt{1.00 - p}}{\sqrt{P}} = \sigma_S = \frac{\sigma_s}{\sqrt{P}}$$
$$\left(\text{where } s = 1 \text{ or } 0 \text{ and } p = \frac{\Sigma s}{P} = S\right) \tag{20}$$

so that where S_i in (14) is a percent $p(1.00 - p)$ should be substituted for σ_i (and σ_j) in (18). In this case the appropriate formula to use for getting r_{ij} in (18) depends on the nature of the distribution of the variable that is expressed in percentage form. If the distribution is normal, tetrachoric r may be appropriate, while if the S in percentage form is from a two point distribution, r from a four fold point surface may be appropriate.

In all the above cases the usual interpretation of the significance of f in respect to sampling errors may be used in entering a normal probability table with a given σ_f from (18) and reading the probability of such a f occurring by chance.[4]

For a numerical illustration of this formula (18), consider the case of two villages, the statistical significance of whose momentums of a social change are to be determined. The data are from a study[1] of Syrian villages where an

[4] Mr. Wilks comments here that, "there is a more exact and rigorous test for comparing the two sets of S's which enter into a pair of M's or F's which involves some recent statistical theory but it is doubtful if the extra refinement is worth while at this stage of sociometric development."

itinerant Health Clinic in two years changed the average hygienic status of the families in each village by amounts of score (on a scale of 1 to 1000 points, devised for this study) as indicated in the table below.

	Village A	Village B
Mean score in 1931 = S_1 =	253	321
" " " 1933 = S_2 =	304	528
Population (families) in 1931 = P_1 =	46	46
" " " 1933 = P_2 =	40	32
Standard deviation of scores in 1931 = σ_1 =	54	39
" " " " " 1933 = σ_2 =	58	70
Families common to both censuses = Q_{12} =	40	32
Correlation of scores from the 2 dates = r_{12} =	.00	.19
$k_{M1} = -(P_1 + P_2)/2Y_{(2-1)} =$	-21.5	-19.5
$k_{M2} = -k_{M1} =$	21.5	19.5
$k_{M1}\sigma_1/P_1 =$	-25.24	-16.53
$k_{M2}\sigma_2/P_2 =$	31.17	42.65
$Q_{12}r_{12} =$	0	6.08
$\sigma_{M'_{2-1}} =$	261	249*
Momentum = M'_{2-1}	1,097	4,037
Significance ratio $M'_{2-1}/\sigma_{M'_{2-1}}$	4.2	16.2

* The calculation of this σ by (18) may be illustrated in detail:

Village B

Coefficients, $\frac{k\sigma}{P}$ → ↓		1	2
		-16.53	42.65
1	-16.53	46 (= P_1) (12,571)	6.08 (= Qr) (−4,286)
2	42.65	6.08 (= Qr) (−4,286)	32 (= P_2) (58,208)

$\Sigma(\ ) = 62{,}207$
$= \sigma^2_{M'(2-1)}$

$\sigma_{M'(2-1)} = 249$

The momentum of the movement towards improved hygiene achieved in village A is 4.2 times its standard error, while that of village B is 16.2 times its standard error. The excess momentum of village A over village B is 8.1 $\left(= \frac{2940}{361}\right)$ times the standard error of their difference in momenta. Since all three of these significance ratios are well over 3 the conclusion is that the observed momenta and difference of momenta are statistically significant and cannot reasonably be due to sampling fluctuations. It may be noted that the significance ratios for the amounts of this social *change*, the difference in mean scores, are in close agreement with the above figures, being 4.1 and 15.9 for

villages A and B respectively, instead of 4.2 and 16.2 as above. These discrepancies of a .1 and .3 in the statistical significance of these social changes compared with the corresponding social momenta are accounted for by the fact that the shift in the size of the population is allowed for in our formula for the case of momenta and is not considered in the usual formula for the case of social change.

A minimum of three measurements of one population is necessary to determine a social force. To determine its standard error all the correlations must be secured between every pair of measurements, each correlation derived from the part of the total population that is common to that pair of measurements. Obviously the data as currently reported from surveys and censuses and statistical bureaus do not meet these specifications. More rigorous analysis of social data and reporting of correlations in it is a prerequisite to the measurement of social forces and their significance.

AMERICAN UNIVERSITY OF BEIRUT, SYRIA.

AN APPROXIMATION TO "STUDENT'S" DISTRIBUTION*

By Walter A. Hendricks

I. Introduction

The function commonly known as "Student's" distribution occupies a prominent position among the classic contributions to the field of statistics, not only for its intrinsic value but also for the stimulus which it gave to statistical research at the time of its discovery.

The function, which may be written in the form,

$$dF_z = \frac{1}{B[\frac{1}{2}(n-1), \frac{1}{2}]}(1+z^2)^{-\frac{1}{2}n}dz, \tag{1}$$

gives the distribution of the ratio, z, of the estimated arithmetic mean, $\bar{x}$, to the estimated standard deviation, s, for samples of n observations drawn from the normal universe specified by the arithmetic mean, zero, and the standard deviation, σ. This function, together with a table of values of its integral was given by "Student."[9, 10]

In view of the fact that similar distributions were subsequently found by Fisher[2] to arise in a larger variety of practical problems than was originally supposed, a table of values of a new integral was later given by "Student"[11] in which the distribution of a variable, t, defined by the relation,

$$t = (n-1)^{\frac{1}{2}}z, \tag{2}$$

rather than the distribution of z itself, was considered. Another table giving the distribution of t, in a form intended to be more convenient for use by research workers wishing to apply statistical methods to experimental data, was later given by Fisher.[3]

The integration of functions of the type defined by equation (1) involves considerable labor, a fact which has been somewhat embarrassing to practical statisticians interested in the distributions of z and t for values of n larger than those included in the above-mentioned tables. The recent appearance of Tables of the Incomplete Beta-Function, prepared under the direction of Pearson,[7] has considerably alleviated the difficulty, but the requirements of certain practical problems are not easily satisfied even with the aid of these tables. Consequently, simple approximations to the distributions of z and t,

---

* A thesis submitted to the Faculty of the Columbian College of The George Washington University in part satisfaction of the requirements for the degree of Master of Arts.

which will be sufficiently accurate for most practical purposes, should be of some interest.

According to "Student,"[9] the distribution of z tends to approach a normal curve with a standard deviation of $(n-3)^{-\frac{1}{2}}$ for values of n greater than 10. However, Deming and Birge[1] have recently suggested that the distribution tends to approach a normal curve with a standard deviation of $(n-1\frac{1}{2})^{-\frac{1}{2}}$.

This thesis presents a simple approximation to the distribution of z, which can be readily extended to the distribution of t and which will give more accurate results than either of the above approximations.

II. Approximation to the Distribution of Z

The approximation presented here is based upon the assumption that, for large values of n, the distribution of s tends to approach a normal curve with the arithmetic mean, $\bar{s}$, and the standard deviation, $\frac{\sigma}{2^{\frac{1}{2}}n^{\frac{1}{2}}}$, that is,

$$dF_s = \frac{n^{\frac{1}{2}}}{\pi^{\frac{1}{2}}\sigma}\, e^{-\frac{n}{\sigma^2}(s-\bar{s})^2}\, ds. \tag{3}$$

Since the distribution of the estimated arithmetic mean, $\bar{x}$, is known to be normal, with the standard deviation, $\frac{\sigma}{n^{\frac{1}{2}}}$, we have for the joint distribution of s and $\bar{x}$:

$$dF_{s,\bar{x}} = \frac{n}{2^{\frac{1}{2}}\pi\sigma^2}\, e^{-\frac{n}{\sigma^2}[\frac{1}{2}\bar{x}^2+(s-\bar{s})^2]}\, ds\, d\bar{x}. \tag{4}$$

$\bar{s}$ may be expressed in terms of n and σ by the well-known relation,

$$\bar{s} = c_n\sigma, \tag{5}$$

in which the factor, c_n, is defined by the formula,

$$c_n = \frac{2^{\frac{1}{2}}}{n^{\frac{1}{2}}}\, \frac{\Gamma(\frac{1}{2}n)}{\Gamma[\frac{1}{2}(n-1)]}. \tag{6}$$

If we write, $c_n\sigma$, in place of $\bar{s}$, in equation (4) and make the transformation,

$$\bar{x} = sz, \tag{7}$$

we have for the joint distribution of s and z:

$$dF_{s,z} = \frac{n}{2^{\frac{1}{2}}\pi\sigma^2}\, e^{-\frac{n}{\sigma^2}[\frac{1}{2}s^2z^2+(s-c_n\sigma)^2]}\, s\, ds\, dz. \tag{8}$$

To find the distribution of z, all that is necessary is to write:

$$dF_z = k\left[\int_{-\infty}^{+\infty} e^{-(as-b)^2}\, s\, ds\right] dz, \tag{9}$$

in which:

$$(10)\qquad \begin{aligned} k &= \frac{n}{2^{\frac{1}{2}}\pi\sigma^2}\, e^{-nc_n^2\frac{z^2}{z^2+2}}, \\ a &= \frac{n^{\frac{1}{2}}}{2^{\frac{1}{2}}\sigma}\,(z^2+2)^{\frac{1}{2}}, \\ b &= 2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\,(z^2+2)^{-\frac{1}{2}}. \end{aligned}$$

The integral in brackets in equation (9) can be evaluated without any difficulty. We have:

$$(11)\qquad \int_{-\infty}^{+\infty} e^{-(as-b)^2}\, s\, ds = \frac{b\pi^{\frac{1}{2}}}{a^2}.$$

Substituting this value in equation (9) and replacing k, a, and b by the quantities which they represent, we obtain the following expression for the distribution of z:

$$(12)\qquad dF_z = \frac{2n^{\frac{1}{2}}c_n}{\pi^{\frac{1}{2}}}\, e^{-nc_n^2\frac{z^2}{z^2+2}}\,(z^2+2)^{-\frac{3}{2}}\, dz.$$

If we now define a new variable, u, by the relation,

$$(13)\qquad u^2 = 2nc_n^2\,\frac{z^2}{z^2+2},$$

and make the appropriate substitutions in equation (12), we have, for the distribution function of u:

$$(14)\qquad dF_u = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}}\, e^{-\frac{1}{2}u^2}\, du.$$

Equation (14) is obviously a normal curve with unit standard deviation. We have thus deduced the interesting fact that, for values of n sufficiently large so that the distribution of s may be represented by a normal curve, the quantity, $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\,\frac{z}{(z^2+2)^{\frac{1}{2}}}$, is distributed as a normal deviate with unit standard deviation.

The accuracy of this approximation as compared with that of the approximation suggested by "Student"[9] and that of the more recent approximation suggested by Deming and Birge[1] may now be considered. As previously stated, the "Student" approximation is based on the assumption that the quantity, $(n-3)^{\frac{1}{2}}z$, is distributed as a normal deviate with unit standard deviation for values of n greater than 10, while that suggested by Deming and Birge is based on the assumption that the quantity, $(n-1\frac{1}{2})^{\frac{1}{2}}z$, is so distributed.

Table 1* gives values of the integral, I_z, defined by:

$$(15)\qquad I_z = \frac{1}{B[\frac{1}{2}(n-1), \frac{1}{2}]}\int_{-\infty}^{z} (1+z^2)^{-\frac{1}{2}n}\, dz,$$

---

* All tables and charts to which reference is made are to be found in the Appendix.

for the case, $n = 10$, together with the corresponding approximate values obtained by making use of the three approximations suggested by "Student," Deming and Birge, and the present author, respectively. The exact values and those obtained by the "Student" approximation were derived from values calculated by "Student"[9] and given by Pearson.[5] All other data in the table were calculated by the present author.

An inspection of Table 1 shows that the values of I_z based on the approximation presented in this thesis agree very well with the corresponding exact values. The agreement is better than that found in the case of either of the other two approximations. The Deming and Birge approximation gives better results than the "Student" approximation for values of z in the neighborhood of zero, but for other values of z the opposite is true.

III. Approximation to the Distribution of t

Since tables giving the distribution of the variable, t, have largely superseded those giving the distribution of z in practical statistical work, the feasibility of applying the above three approximations to the distribution of t is worthy of consideration.

The variable, t, has already been defined in terms of n and z by equation (2). If, in equation (12), we make the transformation,

$$z = (n-1)^{\frac{1}{2}} t, \tag{16}$$

we have, for the distribution function of t:

$$dF_t = \frac{2n^{\frac{1}{2}}(n-1)c_n}{\pi^{\frac{1}{2}}} e^{-nc_n^2 \frac{t^2}{t^2+2(n-1)}} [t^2 + 2(n-1)]^{-\frac{3}{2}}\, dt. \tag{17}$$

If we now define a variable, v, by the relation,

$$v^2 = 2nc_n^2 \frac{t^2}{t^2 + 2(n-1)}, \tag{18}$$

we have, for the distribution function of v:

$$dF_v = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} e^{-\frac{1}{2}v^2}\, dv. \tag{19}$$

Equation (19) shows that, for values of n sufficiently large so that the distribution of s may be represented by a normal curve, the quantity,

$$2^{\frac{1}{2}} n^{\frac{1}{2}} c_n \frac{t}{[t^2 + 2(n-1)]^{\frac{1}{2}}},$$

is distributed as a normal deviate with unit standard deviation. On the other hand, if we assume with "Student" that, for large values of n, the quantity, $(n-3)^{\frac{1}{2}} z$, is normally distributed about zero with unit standard deviation, we should expect to find that the quantity, $\frac{(n-3)^{\frac{1}{2}}}{(n-1)^{\frac{1}{2}}} t$, is also distributed as a normal

deviate with unit standard deviation. If the Deming and Birge approximation to the distribution of z is assumed to be valid, we should expect to find that the quantity, $\frac{(n - 1\frac{1}{2})^{\frac{1}{2}}}{(n-1)^{\frac{1}{2}}} t$, is distributed as a normal deviate with unit standard deviation.

To test the accuracy of each of these three approximations to the distribution of t, we may make use of the well-known table of values of t given by Fisher.[3] This table is so constructed that a value of t corresponding to a given number of "degrees of freedom" and a given value of "P" may be read from the table, where P is defined by the relation,

$$(20) \qquad P = 1 - \frac{2}{(n-1)^{\frac{1}{2}} B[\frac{1}{2}(n-1), \frac{1}{2}]} \int_0^t \left(1 + \frac{t^2}{n-1}\right)^{-\frac{1}{2}n} dt.$$

The entries in the last line of the table, corresponding to an infinite number of "degrees of freedom," are the deviates of a normal curve with unit standard deviation.

To test the accuracy of the "Student" approximation, we may calculate the entries for a line of this table, corresponding to $n - 1$ "degrees of freedom," by multiplying the entries in the last line of the table by $\frac{(n-1)^{\frac{1}{2}}}{(n-3)^{\frac{1}{2}}}$. These approximate values of t may then be compared with the exact values given in the table. The accuracy of the Deming and Birge approximation may be tested in the same manner, except that in this case the entries in the last line of the table should be multiplied by $\frac{(n-1)^{\frac{1}{2}}}{(n-1\frac{1}{2})^{\frac{1}{2}}}$. To test the accuracy of the approximation given by equation (19), we may calculate the values of t corresponding to $n - 1$ "degrees of freedom" by means of the relation,

$$(21) \qquad t^2 = \frac{2(n-1)v^2}{2nc_n^2 - v^2},$$

in which the entries in the last line of the table are to be taken as the values of v.

Table 2 gives the exact values of t corresponding to the values of P given in Fisher's table for $n = 10$, together with the approximate values calculated by means of each of the above three approximations. This comparison of the accuracies of the three approximations is equivalent to the comparisons presented in Table 1. The conclusions which may be drawn are in agreement with those which have already been drawn from that table.

In order to test the behavior of each of the approximations for a larger value of n, values of t corresponding to the different values of P were calculated for $n = 30$. The results are presented in Table 3. The rank of each of the three approximations, with regard to accuracy, for $n = 30$ is the same as for $n = 10$. Although all three give more accurate results for the larger value of n, the superiority of the approximation presented in this thesis is quite apparent.

For extremely large values of n, all three approximations evidently tend to become one-hundred percent accurate, for the distribution of t tends to become normal as n is increased indefinitely. In the case of the "Student" and Deming and Birge approximations, the ratios, $\dfrac{(n-1)^{\frac{1}{2}}}{(n-3)^{\frac{1}{2}}}$ and $\dfrac{(n-1)^{\frac{1}{2}}}{(n-1\frac{1}{2})^{\frac{1}{2}}}$, obviously approach unity, respectively, as n becomes very large. The approximate value of t given by equation (21) also tends to approach the normal deviate, v, as n is increased for we have:

$$\lim_{n\to\infty} t^2 = \lim_{n\to\infty} \frac{2(n-1)v^2}{2nc_n^{\,2} - v^2} = \lim_{n\to\infty}\left[\frac{2nv^2}{2nc_n^{\,2} - v^2} - \frac{2v^2}{2nc_n^{\,2} - v^2}\right]$$
$$(22) \qquad = \lim_{n\to\infty}\left[\frac{v^2}{c_n^{\,2} - \dfrac{v^2}{2n}} - \frac{2v^2}{2nc_n^{\,2} - v^2}\right] = v^2.$$

IV. Discussion

The greater accuracy of the approximation to the distribution of z presented in this thesis apparently can not be explained by the hypothesis that the distribution of s becomes normal more rapidly than the distribution of z as n is increased. Table 4 presents values of the ordinates of the normal curve with unit standard deviation, together with the corresponding ordinates of the exact distributions of the quantities, $\dfrac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$, $(n-3)^{\frac{1}{2}}z$, $(n-1\frac{1}{2})^{\frac{1}{2}}z$, and $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\dfrac{z}{(z^2+2)^{\frac{1}{2}}}$, for $n = 10$. Although the distribution of $\dfrac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$ seems to follow the normal curve more closely than does the distribution of $(n-3)^{\frac{1}{2}}z$, the opposite seems to be true in the case of the distribution of $(n-1\frac{1}{2})^{\frac{1}{2}}z$. The distribution of $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\dfrac{z}{(z^2+2)^{\frac{1}{2}}}$, however, follows the normal curve quite closely.

The behavior of these distributions for $n = 10$ can be observed more easily in Figures 1, 2, and 3 in which the frequency curves of $\dfrac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$, $(n-3)^{\frac{1}{2}}z$, and $(n-1\frac{1}{2})^{\frac{1}{2}}z$ are respectively plotted together with the normal curve with unit standard deviation. The frequency curve of $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\dfrac{z}{(z^2+2)^{\frac{1}{2}}}$ was not plotted because of the fact that this curve follows the normal curve so closely that the two curves could not be distinguished when plotted on the scale used in the other three charts.

The most reasonable conclusion which can be drawn from Table 4 and Figures 1, 2, and 3 is that the departure of the exact distribution of s from the normal curve has very little effect in destroying the normality of the distribution of $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\dfrac{z}{(z^2+2)^{\frac{1}{2}}}$.

V. Values of the Factor, c_n

For the practical application of the approximations to the distributions of z and t presented in this thesis, a table of values of the factor, c_n, is required. Values of this factor, for values of n as high as 100, have been tabulated by Pearson[4, 6] and by Shewhart.[8] For values of n greater than 100, c_n may be calculated accurately to at least five significant figures by the following relation, given by Pearson[4] and by Deming and Birge[1]:

$$c_n = 1 - \frac{3}{4n} - \frac{7}{32n^2}. \tag{23}$$

Table 5 presents values of c_n for some large values of n, calculated by the present author. For values of n not included in this table, c_n may be calculated by means of equation (23) just as rapidly as by interpolation in the table.

VI. Summary and Conclusions

For values of n sufficiently large so that the distribution of s may be represented by a normal curve, the quantities,

$$2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2+2)^{\frac{1}{2}}} \text{ and } 2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{t}{[t^2+2(n-1)]^{\frac{1}{2}}},$$

are distributed as normal deviates with unit standard deviation. The results obtained by assuming a normal distribution of s are more accurate than those obtained by assuming that either $(n - 3)^{\frac{1}{2}}z$ or $(n - 1\frac{1}{2})^{\frac{1}{2}}z$ is distributed as a normal deviate with unit standard deviation. For extremely large values of n, the distribution of each of the above quantities tends to approach a normal curve with a mean of zero and unit standard deviation.

VII. References

(1) Deming, W. Edwards, and R. T. Birge, 1934. On the statistical theory of errors. Reviews of Modern Physics, 6: 119–161.

(2) Fisher, R. A., 1925. Applications of "Student's" distribution. Metron, 5: 90–104.

(3) Fisher, R. A., 1934. Statistical Methods for Research Workers, 5th ed. Oliver and Boyd, Edinburgh and London.

(4) Pearson, Karl, 1915. On the distribution of the standard deviations of small samples. Biometrika, 10: 522–529.

(5) Pearson, Karl, 1924. Tables For Statisticians And Biometricians, Part I, 2nd ed. Cambridge University Press, Cambridge.

(6) Pearson, Karl, 1931. Tables For Statisticians And Biometricians, Part II, 2nd ed. Cambridge University Press, Cambridge.

(7) Pearson, Karl, 1934. Tables Of The Incomplete Beta-Function. Cambridge University Press, Cambridge.

(8) Shewhart, W. A., 1931. Economic Control Of Quality Of Manufactured Product. D. Van Nostrand Co., New York.

(9) "Student," 1908. The probable error of a mean. Biometrika, 6: 1–25.

(10) "Student," 1917. Tables for estimating the probability that the mean of a unique sample of observations lies between $-\infty$ and any given distance of the mean of the population from which the sample is drawn. Biometrika, 11: 414–417.

(11) "Student," 1925. New tables for testing the significance of observations. Metron, 5: 105–120.

THE GEORGE WASHINGTON UNIVERSITY.

VIII. Appendix

TABLE 1

Exact values of I_z and approximate values, derived from tables of the normal probability integral, for $n = 10$

z	I_z			
	Exact value	"Student" approximation	Deming & Birge approximation	Hendricks approximation
−2.0	.0001	.0000	.0000	.0004
−1.8	.0002	.0000	.0000	.0006
−1.6	.0005	.0000	.0000	.0010
−1.4	.0011	.0001	.0000	.0018
−1.2	.0029	.0007	.0002	.0038
−1.0	.0075	.0041	.0018	.0086
− .8	.0199	.0171	.0098	.0211
− .6	.0527	.0562	.0401	.0535
− .4	.1304	.1448	.1218	.1307
− .2	.2816	.2984	.2799	.2817
.0	.5000	.5000	.5000	.5000
+ .2	.7184	.7016	.7201	.7183
+ .4	.8696	.8552	.8782	.8693
+ .6	.9473	.9438	.9599	.9465
+ .8	.9801	.9829	.9902	.9789
+1.0	.9925	.9959	.9982	.9914
+1.2	.9971	.9993	.9998	.9962
+1.4	.9989	.9999	1.0000	.9982
+1.6	.9995	1.0000	1.0000	.9990
+1.8	.9998	1.0000	1.0000	.9994
+2.0	.9999	1.0000	1.0000	.9996

TABLE 2

Exact values of t corresponding to different values of P and approximate values, derived from normal deviates, for $n = 10$

P	t			
	Exact value	"Student" approximation	Deming & Birge approximation	Hendricks approximation
.90	.129	.142	.129	.129
.80	.261	.287	.261	.261
.70	.398	.437	.396	.398
.60	.543	.595	.540	.544
.50	.703	.765	.694	.703
.40	.883	.954	.866	.884
.30	1.100	1.175	1.066	1.104
.20	1.383	1.453	1.319	1.386
.10	1.833	1.865	1.693	1.844
.05	2.262	2.222	2.017	2.290
.02	2.821	2.638	2.394	2.896
.01	3.250	2.921	2.650	3.389

TABLE 3

Exact values of t corresponding to different values of P and approximate values, derived from normal deviates, for $n = 30$

P	t			
	Exact value	"Student" approximation	Deming & Birge approximation	Hendricks approximation
.90	.127	.130	.127	.127
.80	.256	.263	.256	.256
.70	.389	.399	.389	.389
.60	.530	.543	.529	.530
.50	.683	.699	.680	.683
.40	.854	.872	.849	.854
.30	1.055	1.074	1.045	1.055
.20	1.311	1.328	1.293	1.312
.10	1.699	1.705	1.659	1.700
.05	2.045	2.031	1.977	2.047
.02	2.462	2.411	2.347	2.466
.01	2.756	2.670	2.598	2.764

TABLE 4

Ordinates of the normal curve with unit standard deviation and ordinates of the exact distribution functions of $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$, $(n-3)^{\frac{1}{2}}z$, $(n - 1\frac{1}{2})^{\frac{1}{2}}z$, *and* $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2+2)^{\frac{1}{2}}}$ *for* $n = 10$

Deviation from mean	Ordinates of distribution function				
	Normal deviate	$\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$	$(n-3)^{\frac{1}{2}}z$	$(n - 1\frac{1}{2})^{\frac{1}{2}}z$	$2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2+2)^{\frac{1}{2}}}$
−3.0	.0044	.0006	.0071	.0108	.0034
−2.5	.0175	.0085	.0181	.0254	.0156
−2.0	.0540	.0454	.0459	.0581	.0544
−1.5	.1295	.1356	.1092	.1234	.1306
−1.0	.2420	.2663	.2256	.2290	.2426
− .5	.3521	.3751	.3692	.3454	.3522
.0	.3989	.3999	.4400	.3991	.3990
+ .5	.3521	.3343	.3692	.3454	.3522
+1.0	.2420	.2245	.2256	.2290	.2426
+1.5	.1295	.1233	.1092	.1234	.1306
+2.0	.0540	.0560	.0459	.0581	.0544
+2.5	.0175	.0213	.0181	.0254	.0156
+3.0	.0044	.0068	.0071	.0108	.0034

TABLE 5

Values of c_n *for large values of* n

n	c_n	n	c_n
100	.99248	900	.99917
150	.99499	1000	.99925
200	.99624	2000	.99962
250	.99700	3000	.99975
300	.99750	4000	.99981
350	.99786	5000	.99985
400	.99812	10000	.99992
450	.99833	20000	.99996
500	.99850	30000	.99997
600	.99875	40000	.99998
700	.99893	50000	.99998
800	.99906	100000	.99999

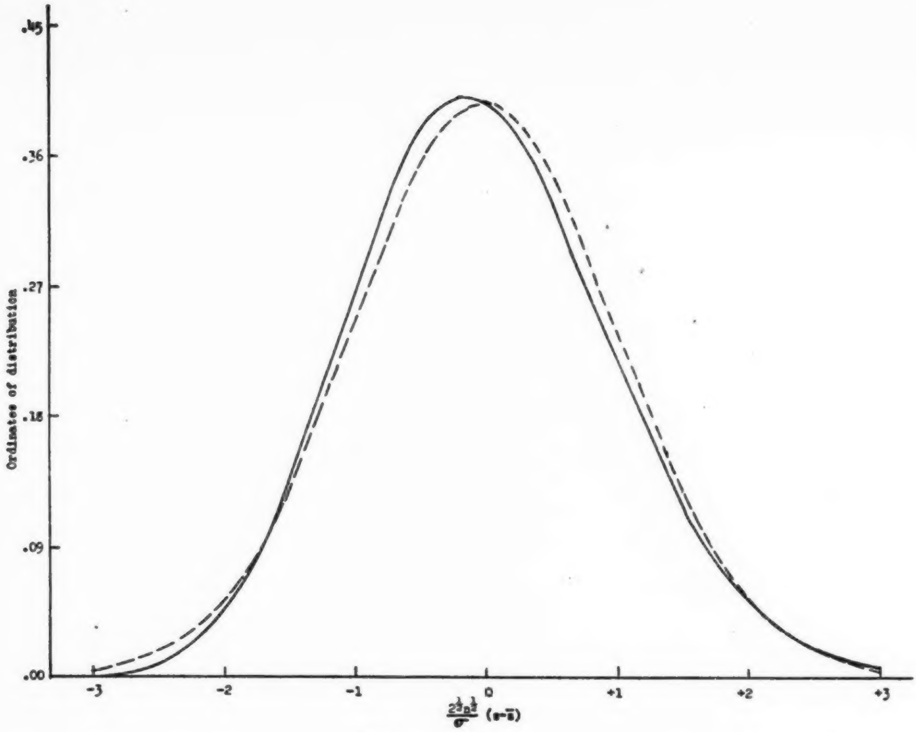

FIG. 1. EXACT DISTRIBUTION OF $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$ FOR $n = 10$ AND NORMAL CURVE WITH UNIT STANDARD DEVIATION

———, exact distribution; ----------, normal curve

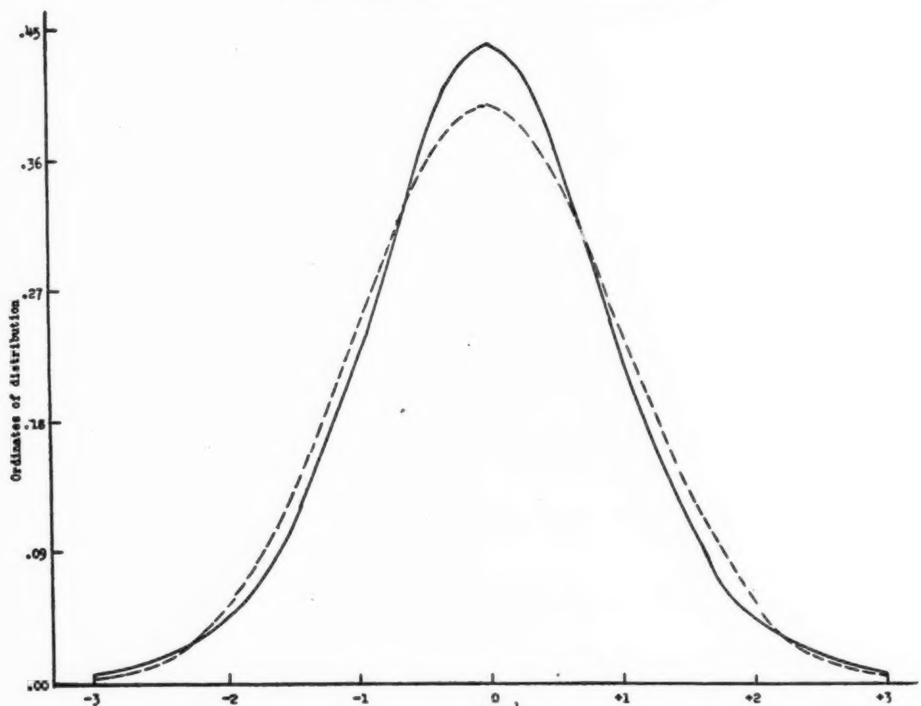

FIG. 2. EXACT DISTRIBUTION OF $(n - 3)^{\frac{1}{2}}z$ FOR $n = 10$ AND NORMAL CURVE WITH UNIT STANDARD DEVIATION

———, exact distribution; ----------, normal curve

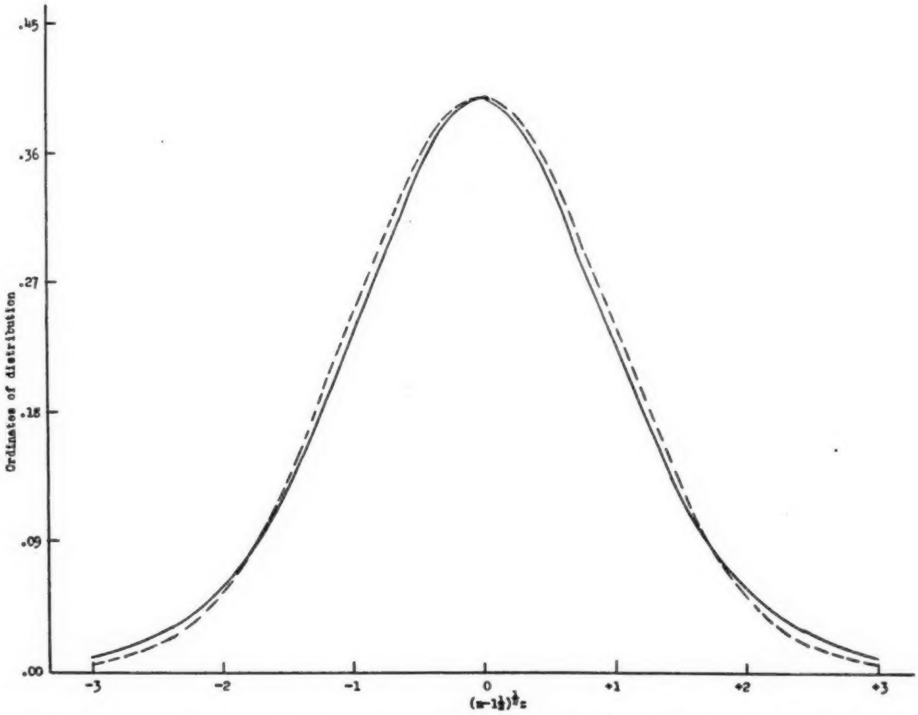

FIG. 3. EXACT DISTRIBUTION OF $(n - 1\frac{1}{2})^{\frac{1}{2}}z$ FOR $n = 10$ AND NORMAL CURVE WITH UNIT STANDARD DEVIATION

———, exact distribution; ----------, normal curve

TABLE OF CONTENTS